Session 2. Quality and Confidentiality of data / Data science techniques for the collection and analysis of financial information

Stefan Bender, Head of Research Data and Service Center (RDSC), Deutsche Bundesbank

Financial Information Forum of Latin American and the Caribbean Central Banks
VI Meeting – virtual format
Mexico City, 27 - 29 May 2020

Based on a joint project with and contributions from:

(The views expressed here do not necessarily reflect the opinion of the Deutsche Bundesbank or the Eurosystem.)
Café Terrace at Night is an 1888 oil painting by the Dutch artist Vincent van Gogh. Van Gogh painted Café Terrace at Night in Arles, France.
Café Terrace at Night is an 1888 oil painting by the Dutch artist Vincent van Gogh. Van Gogh painted Café Terrace at Night in Arles, France.
Overview

- (Motivation)

- **Data Science for the Collection of Financial Information**
  - The Research Data and Service Center of Bundesbank (RDSC)
  - New Sources and Techniques
  - Reproducibility
  - FAIR Data, Annodata Schema

- **Data Science for the Analysis of Financial Information**
  - Enhance Data Services for Users
  - Justify added Value of Microdata for Research (and beyond)

- **Conclusion**

  THX to the INEXDA members for their discussions and work on some of the topics, which will be presented!
Motivation

- **Aggregate datasets** are important for monitoring macroeconomic developments and macroeconomic policy.

- **Granular data** is necessary to understand global developments and in particular differences across countries.

- Combining datasets and looking beyond aggregate statistics into heterogeneous developments require the transformation of “data“ into “knowledge“.

- **Local constraints** make it difficult, or often impossible, to link micro datasets from different jurisdictions, even for research and financial stability analysis.

- Better accessibility and sharing of granular data would open up new possibilities for analysis by providing new insights into the effect of policies.

What can we do from the statistical side to support this process?
What’s new in (central bank) statistics?

- Micro data overhaul the traditional value-added chain in central banking statistics.
  - Traditional central banking statistics are collected for a **specific purpose**.
  - Micro data are collected only once and can be used for **multiple purposes**: The statistical reporting burden declines.
  - **Data protection** becomes more challenging.

- **Technological innovations** have revolutionized the infrastructure for collecting, storing, and using micro-data.
  - Advanced knowledge in storage and organization of large (integrated) micro-data.
  - Improved tools for analyzing and processing microdata.
  - Cheaper storage technologies.
  - Standardization.

- Official statistics has **lost the monopoly** in providing information to society.
Tasks of the Research Data and Service Center (RDSC)

The RDSC offers access for **non-commercial** research to (highly sensitive) micro data of the Bundesbank. Microdata for banks, companies, securities and households are available:

- Generate (linked) micro data
- Offer advisory service on data selection and data access (data handling, research potential, scope and validity of data)
- Provide data access and data protection
- Document data and methodological aspects of the data
- Work on own research projects (in close cooperation with the Bank’s business areas and the Research Centre)
- Organize conferences and workshops.
Factsheet on the RDSC

- 20 employees
- 12 working places for guest researchers in Frankfurt (fully booked several times).
- 2 working places in Düsseldorf

In 2018:
- Around 130 project applications, 73 were realized
- Over 2,000 files (over 3.5 million lines) checked (output control)
- Average of used data products per research project: 2.68
- Papers of RDSC users are out

In 2017 over 300 active projects, over 160 institutions involved (around 90 non-German).
Information Life Cycle of the RDSC (but could be transformed)
Information Life Cycle of the RDSC: Part 1

Research Data Centre / Statistic Department

Research

Data Service
Available microdata at the RDSC

- Banks
- Companies
- Securities
- Households
Coming back to the Dutch Painting
New Data Sources

Bundesbank is collecting the needed data (for example by regulations), but for fulfilling the different purposes, Bundesbank is using more and more “additional data sources”:

- Internet platform information like house prices
- Truck toll mileage (Destatis)
- Patent data
- Commercial Data
- Expectation Surveys, which will be combined with other data (informed consent)
- Unstructured Data (RDSC is offering speeches of the Governor Board)
Securities Holdings Statistics

- German banks provide monthly reports of securities holdings (security-by-security)
- DQM with labor intensive manual case-by-case evaluations

Goal:
- Support compilers by providing predictions of the result of manual checks with machine learning methods

Experience:
- 50% reduction in time for check and
- increased effectiveness of evaluations

Prediction allows for sorting
Descriptive Analysis of Patterns
Number of Reporting Banks, Days since Maturity (Tobias Cagala, S5)
Company data (non financial institutions (NFI)):

There is no common unique firm identifier in Germany.
   (Company business register-ID not stable)

We have to match firm data…
  ▪ … that do not have a common unique identifier / key
  ▪ … by using alternative identifiers (such as names, addresses, sectors, legal forms)

RDSC has matched several NFI-microdatasets (from Statistics, Banking Supervision and external data) with an advanced machine learning algorithm and generated a matching table (with probabilistic matching scores)

Goal:
  ▪ Improve data quality, increase analytical value of data
  ▪ More general and flexible Record Linkage System
  ▪ Historicized matching tables
Evaluation: Record Linkage of NFI-Microdata (Christopher-Johannes Schild)

Precision = \( \frac{TP}{TP + FP} = 98.0\% \)

Recall / Coverage = \( \frac{TP}{TP + FN} = 93.3\% \)
Data Generating Process

Until now (in many cases): ad hoc generation of data for research.

RDSC has started to/with:
- Establishing standardised data products.
- Implementing RDSC data quality procedures.
- Documentation of data.
- Harmonisation of data.
- Register data to get data identifiers (DOIs).
Information Life Cycle of the RDSC: Part 2

Public

Publications

Research Data Centre

Research

Data Service
The 5 Safes in the RDSC

- **Safe people**: non-disclosure agreement, contract (with penalty up to 60,000 Euro, publishing the name, exclusion from access up to 2 years).
- **Safe projects**: non-commercial research, project description.
- **Safe environment**: working places without internet connection, (cell) phone, photo, printer and drive.
- **Safe data**: (weakly) anonymized data.
- **Safe results**: output control, papers/presentations are checked.

**Access to real data**, anonymization is only one dimension, others have more effects on data protection.
A „New“ Challenge: Reproducibility

Financial Information Forum of Latin America and the Caribbean Central Banks - VI Meeting (virtual format)
27 - 29 May 2020
Page 21
FAIR Data and the Annodata Concept

Public

FAIR
- Findable
- Accessible
- Interoperable
- Reproducible

Annodata Concept

Research Data Centre

Publications

Research

User specific knowledge

Data Service

Financial Information Forum of Latin America and the Caribbean Central Banks - VI Meeting (virtual format)
27 - 29 May 2020
Page 23
Access regime

Database

Dataset family

Record linkage

Combining restrictions

Global rules

1-3 on dataset family level
4-6 on global level (i.e. irrespective of researcher affiliation, research field, and access mode)
7-8 on project level
Summing Up the First Part

Public

FAIR
- Findable
- Accessible
- Interoperable
- Reproducible

User specific knowledge

Research Data Centre

Output (Results)

Publication.do

Master.do

Standard rules for programming

Data Service

Original Data

Financial Information Forum of Latin America and the Caribbean Central Banks - VI Meeting (virtual format)
27 - 29 May 2020
Page 25
Part 2

- Data Science for the Analysis of Financial Information
  - Enhance Data Services for Users
  - Justify added Value of Microdata for Research (and beyond)
Information Life Cycle of the RDSC (but could be transformed)
Motivation

1. Enhance data services for users
   - Build Amazon-style recommendation system
   - *What have others done with the data?*

2. Justify added value of microdata for research
   - Dataset impact factor
1. Worldwide competition* with participants from all over the world and with New York University in the lead
   - Meaningful solutions:
     - Allen AI
     - KAIST
     - GESIS
     - Paderborn University
     - First step for a systematic approach to find data sets in publications.

2. Contact with RePEc who think about including dataset mentions into their system


* For more information see https://coleridgeinitiative.org/richcontextcompetition/workshopagenda
3 Data Description

The data sources used in our study are (i) Auxmoney for data on P2P lending; (ii) the Deutsche Bundesbank (Interest Rates Statistics) for data on bank lending; (iii) Schufa for data on credit ratings; (iv) the Deutsche Bundesbank (Balance Sheet Statistics) for data on loan loss provisions.

Auxmoney is the oldest and largest P2P lending platform in Germany. According to its website, from the day it began business in 2007 until late 2015, the total volume of credit provided was €219 million in 39,000 projects, with an average nominal interest rate of 9.65%.

Auxmoney provided us with two different datasets. The first includes all loans divided by state between January 2010 and September 2014, with no maturity information. The second includes the average interest rate and the average credit rating represented by the Schufa score for each state per month.

The Deutsche Bundesbank statistics used in this study are provided by two different datasets. The first is the Interest Rates Statistics (MIR, see Baade and Beier (2016) for further information on this data source), which is a stratified sample of the German banking sector used for supervisory activities and gives the amounts and the interest rates per bank and per month applied to nonconstruction consumer credit lines (outstanding and new business) for different maturities (overdraft, up to one year, and more than one year). The statistics are composed of monthly observations between January 2010 and September 2014. The second is the dataset from the Balance Sheet Statistics (BiSTA, see Beier, Krüger, and Schneider (2016) for further information on this data source), which gives information on write-ups and write-downs, from which we derive the banks’ loan loss provisions.

Our analysis is at the bank-state level. The regional differentiation of bank loans is possible because of a feature of the German banking system: the presence of Sparkassen (savings banks) and Volksbanken (cooperative banks). Each bank is only present in one German state. Sparkassen are geographically restricted banks with a legal mandate to provide bank services to all creditworthy

\[^{19}\text{Schufa is a German private credit bureau with 479 million records on 86.2 million natural persons. Schufa provides credit ratings for each person requesting a loan and Auxmoney provides the Schufa score of each credit application.}
^{20}\text{For reasons of data confidentiality, Auxmoney provides its credit intermediation by month and state only if five or more loans were made in that month in that state.}
^{21}\text{The Interest Rates Statistics (MIR) is the German part of a larger dataset that is used by the ECB for regulatory purposes. It does not cover the whole German banking sector, only a stratified sample. For this reason, our dataset only covers the banks in the dataset.}

Getting the data: Scraping central bank publications

RePEc

Economics Departments, Institutes and Research Centers in the World
Central Banks, Monetary Authorities

Source: https://edirc.repec.org/central.html

➤ Over 57,000 central bank publications downloaded

Thanks to GALILEO-Team

Financial Information Forum of Latin America and the Caribbean Central Banks - VI Meeting (virtual format)
27 - 29 May 2020
Model evaluation with central bank sample

Two algorithms to find data sets in publications:

- Allen AI
- KAIST

Experiences:

- Allen AI only finds what it is trained on
- KAIST also finds unknown datasets
- Yet, KAIST is very noisy
<table>
<thead>
<tr>
<th>Title</th>
<th>KAIST</th>
<th>Allen AI</th>
</tr>
</thead>
<tbody>
<tr>
<td>The PHF: A comprehensive panel survey on household finances and wealth in Germany</td>
<td>PHF</td>
<td>PHF, Panel on Household Finances</td>
</tr>
<tr>
<td>Saving and learning: Theory and evidence from saving for child's college</td>
<td>PSID s family data, PSID, Panel Study of Income Dynamics, Data, rst surveyed in the spring of, CDS TA</td>
<td>Panel Study of Income Dynamics (PSID)</td>
</tr>
<tr>
<td>Monetary policy and the oil futures market</td>
<td>commodity prices, VAR, JEL Classification</td>
<td>- / -</td>
</tr>
<tr>
<td>Is the willingness to take financial risk a sex-linked trait? Evidence from national surveys of household finance</td>
<td>DNB Household Survey, BIS Papers</td>
<td>- / -</td>
</tr>
<tr>
<td>China's role in global inflation dynamics</td>
<td>WEO, UN Comtrade database, Journal of Econometrics, OECD s Main Economic Indicators</td>
<td>- / -</td>
</tr>
</tbody>
</table>
From “unrelated” articles to data network

Example: BBk’s monthly Balance Sheet Statistics (BISTA)

1 BBk credit register

2 Data on securities
From dataset network to recommendations

Related to data you've viewed:

1. Jointly used datasets
2. Publications
3. Similar data
4. …

Thanks to Julia Lane
From dataset network to dataset impact factor

Number of publications per dataset and JEL code

Source: Own calculations
Deutsche Bundesbank

Financial Information Forum of Latin America and the Caribbean Central Banks - VI Meeting (virtual format)
27 - 29 May 2020
Page 36
Dataset impact factor: Our axioms

✓ More publications using dataset are better
✓ Broader usage is better
  ▪ More links to other datasets are better
  ▪ More recent usage is better
  ▪ Higher impact factor of publications is better
  ▪ Dataset is better when it sparks new research
Implications

Evidence-based policy

- Micro Data
- Research Publications
  - Policy briefs
  - Newspaper articles
  - Speeches
  - Formal consultations

- Measure data impact on political decisions
- Better allocate resources to improve important data
- Use data efficiently through data recommender
User interface to explore micro data usage
Cyclicality of SME lending and government involvement in banks

Journal of Banking & Finance, 77, 64-77, April 2017
https://doi.org/10.1016/j.jbankfin.2017.01.010

Authors
Patrick Behr - Fundação Getúlio Vargas
Daniel Fuchs - German Federal Bank
Lars Norden - Fundação Getúlio Vargas
Corresponding Author

Abstract
Recent regulatory efforts aim at lowering the cyclicality of bank lending because of its potentially detrimental effects on financial stability and the real economy. We investigate the cyclicality of SME lending of local banks with versus without a public mandate, controlling for location, size, loan maturity, capitalization, funding structure, liquidity, profitability, and credit demand-side factors. The public mandate is set by local governments and stipulates a sustainable provision of financial services to local customers and a deviation from strict profit maximization. We find that banks with a public mandate are 25% less cyclical than other local banks. The result is credit supply-side driven and especially strong for public mandate banks with high liquidity and stable deposit funding. Our findings have implications for the bank structure, financial stability and the finance-growth nexus in a local context.

Dataset references - 2

Gewinn- und Verlustrechnung der Banken (GUV)
Monatliche Bilanzstatistik (BISTA)

Publication references - 25

Social Capital and the Viability of Stakeholder-Oriented Firms: Evidence from Savings Banks*
Charlotte Ostergaard, Iboya Schindel, Bent Vale
2016, Review of Finance - Article

Small Banks and Local Economic Development*

Financial Information Forum of Latin America and the Caribbean Central Banks - VI Meeting (virtual format)
27 - 29 May 2020
Get dataset information and inspiration

Data Set - Gewinn- und Verlustrechnung der Banken (GUV)

https://doi.org/10.12757/dBk.Guv.9516.01

Abstract


More

Financial Information Forum of Latin America and the Caribbean Central Banks - VI Meeting (virtual format)
27 - 29 May 2020
Page 41
Summing Up I: General

- **Developments** for RDSC(s) and INEXDA was/are fast, but **incremental**: trust building, growing data complexity, learning process …
- (New) **skills** for researchers / data producers.
- **Engagement** of researchers (value of data work, data impact factor).
- **Efficiency**: access system in a RDC, metadata/recommendation system project management in a RDC, …
- **Knowledge exchange**: Financial Big Data Cluster, Tech Campus, GAIA-X.
- **Harmonization/Internationalization**: G20 initiative on data sharing and data access of central banks, INEXDA.
Measure data impact on political decisions.

Use data efficiently through data recommender (from ego-network).

Collaboration
- Knowledge sharing
- Metadata

Secure workspace
- Services and Tools

Better allocate resources to improve data quality and service.

Data Stewardship
- Approval
- Monitoring
- Reporting

*Example: Joint dataset usage in publications for the BBk’s monthly Balance Sheet Statistics (BISTA)
Thank you!

- **Website**: www.bundesbank.de/fdsz
- **Contact**: fdsz@bundesbank.de