Experiences of the Deutsche Bundesbank
Session 3. Financial information infrastructures based on microdata
Prof. Stefan Bender, Research Data and Service Center, Deutsche Bundesbank

Financial Information Forum of Latin American and the Caribbean Central Banks
V Meeting
28 and 29 May 2019
Lima, Peru

The views expressed here do not necessarily reflect the opinion of the Deutsche Bundesbank or the Eurosystem.
Overview

- Microdata at the Bundesbank
- IMDIAS
- Knowledge Life Cycle (IDIS-R)
- Use of „New Techniques“
  - Data Quality
  - Record Linkage
- Rich Context
- Conclusion
Available microdata at the RDSC

- Banks
- Companies
- Securities
- Households
Bundesbank’s relevant microdata sources and their connections (excerpt)

28 May 2019, Lima, Peru
Stefan Bender, Research Data and Service Center, Deutsche Bundesbank
Policy evaluation can make better use of existing datasets

- The Bundesbank – like other central banks – produces datasets which are highly valuable for policy analysis and research.
  - So far, most of these datasets have been used to provide aggregate statistics and ad hoc analysis of specific policy issues.
  - There is significant knowledge of data and institutional background.

- **Systematic use of these data for policy analysis is often constrained by**
  - Time
  - IT-resources
  - Legal restrictions

- **The Bundesbank has launched a large-scale initiative aimed at making better use of existing data both, for policy analysis as well as internal and external researchers.**
Granular data become more and more important for assessing monetary and regulatory policy as well as for financial stability. Hence the Bundesbank has launched the large-scale initiative IMIDIAS aimed at making better use of existing data both, for policy analysis as well as internal and external researchers.

**Goals of IMIDIAS:**

- Support policymaking process
- Encourage cooperation with (external) researchers
- Promote evidence-based policy-making
Knowledge Life Cycle (IDIS-R)
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

<table>
<thead>
<tr>
<th>isin</th>
<th>prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>DE000...</td>
<td>82</td>
</tr>
<tr>
<td>DE000...</td>
<td>81</td>
</tr>
<tr>
<td>DE000...</td>
<td>80</td>
</tr>
<tr>
<td>DE000...</td>
<td>80</td>
</tr>
<tr>
<td>DE000...</td>
<td>73</td>
</tr>
<tr>
<td>DE000...</td>
<td>60</td>
</tr>
<tr>
<td>DE000...</td>
<td>43</td>
</tr>
<tr>
<td>DE000...</td>
<td>42</td>
</tr>
<tr>
<td>DE000...</td>
<td>39</td>
</tr>
<tr>
<td>DE000...</td>
<td>35</td>
</tr>
<tr>
<td>DE000...</td>
<td>2</td>
</tr>
<tr>
<td>DE000...</td>
<td>2</td>
</tr>
<tr>
<td>DE000...</td>
<td>1</td>
</tr>
<tr>
<td>DE000...</td>
<td>1</td>
</tr>
</tbody>
</table>
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 = $\frac{TP}{TP + FN} = 93.3\%$

Precision = $\frac{TP}{TP + FP} = 98.0\%$
Recall / Coverage = $\frac{TP}{TP + FN} = 93.3\%$
Building a dataset recommendation engine
Experiences from a machine learning competition in text mining

Hendrik Doll, Research Data and Service Centre (RDSC)

Based on a project with and contributions from:
Stefan Bender, Christian Hirsch, Julia-Katharina Ginz, John Chase,
Jonathan Morgan, Ian Mulvaney, Andrew Gordon and Julia Lane

The views expressed in this presentation do not necessarily reflect the opinion of the Deutsche Bundesbank or the Eurosystem.
Motivation

Status Quo: Linear information flow
- Researchers use microdata and metadata to produce outcome (publications)
- No structured knowledge flows back to data producers and data documentation

Goal: Enable feedback loops
- Make knowledge generated in the research process usable
- Two sources of information to extract:
  - Tacit knowledge: Enable and incentivize data users to feedback information
  - Automatic: Text mining to find information in research publications
Framework: Knowledge Life Cycle

- **Data Stewardship**
  - Approval
  - Monitoring
  - Reporting

- **Collaboration**
  - Knowledge sharing
  - Metadata

- **Secure workspace**
  - Services and Tools

- **Rich Context**
  - Collaboration
  - Knowledge sharing
  - Metadata
  - Secure workspace
  - Services and Tools

- **User Specific Knowledge**
  - Publications
  - Data Service
  - Research

Stefan Bender, Research Data and Service Center, Deutsche Bundesbank
28 May 2019, Lima, Peru
1. What is the added value of microdata for research?
   - What is added value of Bundesbank data? (BISTA vs. Bankscope)
   - Dataset impact factor (Research data is a public good, show outcome to justify societal investment)

2. Enhance data services for user
   - Strengthen quantitative research through optimal microdata usage
   - Design of Amazon-type proposition system for researchers („What data is out there?“ „What have others done with the data?“)
Related to data you've viewed

New data similar to data you've used

What others have done with similar data (recipes)

Recipes like yours
So how do we do this?

- **Step 1:** Create the set of corpora and metadata (computer science technology) - Competition
  
- **Step 2:** Figure out how you learn from it and automate it (machine learning techniques) – Engagement

- **Step 3:** Gamification – recognize and emphasize patterns (with human curation) – Rinse and repeat

THX to Hendrik Doll and Christian Hirsch
Run competition
(https://coleridgeinitiative.org/richcontextcompetition)

Rich Context Competition
Workshop Agenda

Summary
Join us to hear presentations from the finalists for the NYU Coleridge Initiative’s Rich Context Competition. The competition challenged computer scientists to find ways of automating the discovery of research datasets, fields and methods used in social science research publications.

The teams are representatives from GESIS, KAIST, Paderborn University, and Allen AI

THX to Hendrik Doll and Christian Hirsch
Hand-curate corpus if needed - example

- Training dataset: 5,000 labeled publications with dataset usage, fields and methods

<table>
<thead>
<tr>
<th>Publication Title</th>
<th>Authors</th>
<th>Data set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clustering or competition? The foreign investment behaviour of German banks</td>
<td>Lipponer A., Buch C.</td>
<td>MIDI, BISTA, GUV</td>
</tr>
<tr>
<td>How will Basel II affect bank lending to emerging markets? An analysis based on</td>
<td>Liebig T., Porath D.,</td>
<td>MIMIK</td>
</tr>
<tr>
<td>German bank level data</td>
<td>di Mauro B., Wedow M.</td>
<td></td>
</tr>
<tr>
<td>FDI versus cross-border financial services: The globalisation of German banks</td>
<td>Buch C., Lipponer A.</td>
<td>MIDI, BISTA, GUV</td>
</tr>
<tr>
<td>German bank lending during emerging market crises: A bank level analysis</td>
<td>Heid F., Nestmann T.,</td>
<td>MIMIK</td>
</tr>
<tr>
<td></td>
<td>di Mauro B., Westernhagen N.</td>
<td></td>
</tr>
</tbody>
</table>
Step 1: competition design

Datasets

Topics

Methods

THX to Hendrik Doll and Christian Hirsch

Stefan Bender, Research Data and Service Center, Deutsche Bundesbank
28 May 2019, Lima, Peru

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 Bode 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 Schneer (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 individuals.

1Schufa is a German private credit bureau with 473 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.

2For 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.

The Interest Rates Statistics (MIr) is the German part of a larger dataset that is used by the ECE for regulatory purposes. It does not cover the whole German banking sector, only a stratified sample. For this reason, our sample does not cover all Sparkassen and Volksbanken in Germany, just the ones present in this data source.

P2P Lenders versus Banks:
Cream Skimming or Bottom Fishing?

Céleste de Roovere – Loriane Pettizer – Arjana Fisker
SAFE Working Paper No. 104
Application: Design of Amazon-type proposition system for researchers
Where we currently are

Retrieving new unlabeled corpus

- Extending the corpus to broader finance and economics literature
- Special focus: Central bank publications

Text mining or potentially labelling by external services

- Testing the models with new corpus
- Potentially and complementary: Outsourcing the labelling to external provider

Developing analytical tools/applications

Creating prototypes for data set recommendation system and data set impact factor
Take-Aways

WHAT WE DID

- Close the gap between publications and data services
- Engage users to contribute knowledge to existing metadata

WHAT’S NEXT

- Digital integrated system around data-services (as a single point of reference for users)
Getting the Knowledge Life Cycle …
... into the next level

[Diagram showing the relationships between Publications, Data Service, Research, User Specific Knowledge, and a new level]

Stefan Bender, Research Data and Service Center, Deutsche Bundesbank
28 May 2019, Lima, Peru
Thank you!

- **Website:** www.bundesbank.de/fdsz
- **Contact:** fdsz@bundesbank.de
Other Applications of Bundesbank in Machine Learning

- **Saisonal Adjustment**

- **Identification of Holdings**
  - by Frank Raulf

- **Record Linkage**
  - *(we will see in some minutes)*