

WORKSHOP ON DATA SECURITY AND PRIVACY
BANCO DE PORTUGAL AND IRVING FISHER COMMITTEE
14 AND 15 SEPTEMBER 2026, LISBON

EXPLORING PRIVACY-PRESERVING METHODS FOR LARGE-SCALE CREDIT MICRODATA AT BELAB

EUGENIA KOBLENTS, STATISTICS DEPARTMENT
DATA SCIENCE EXPERT

14 SEPTEMBER 2026

BANCODE **ESPAÑA**
Eurosistema



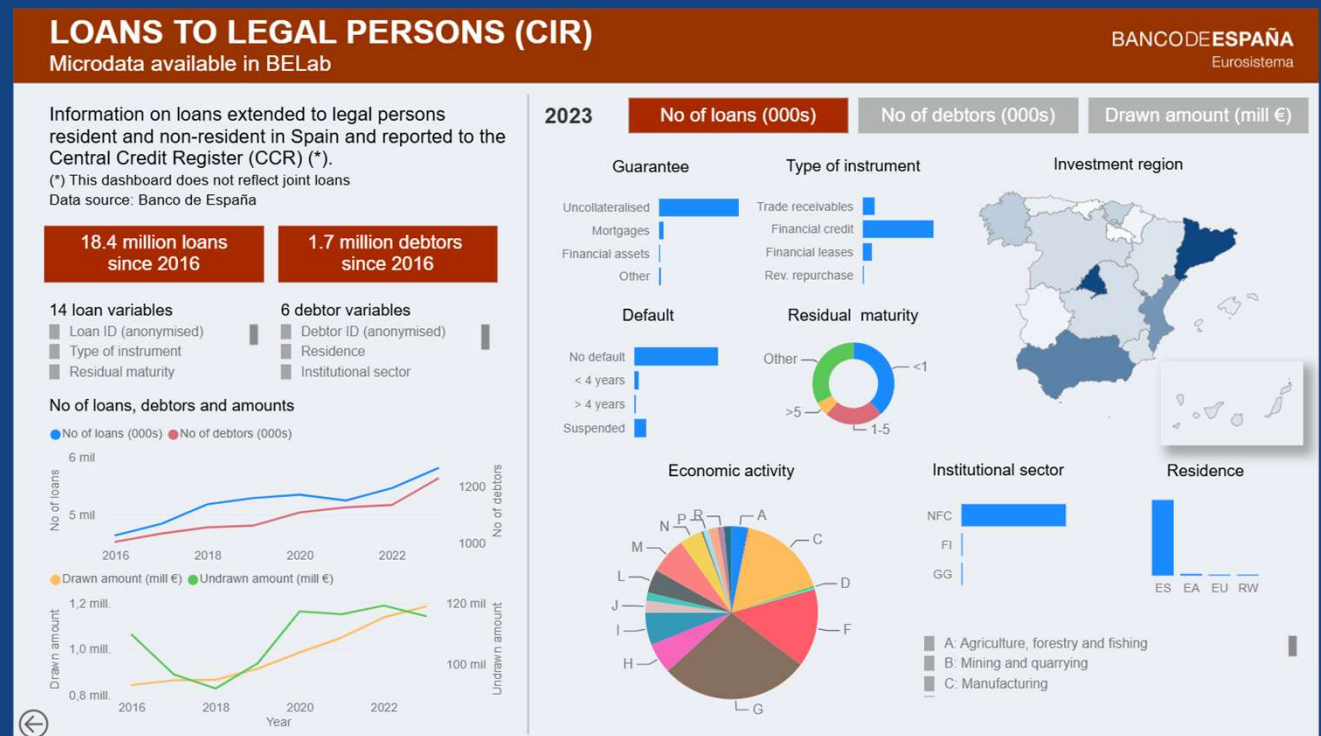


ÍNDICE

1. Motivation and background
2. Secondary anonymisation of CIR and CIR-CBI datasets
3. Secondary anonymisation of the large-scale CIR-PF dataset:
 - Alternative methods
 - Selected approach
4. Conclusions and future work

MOTIVATION AND BACKGROUND

- ❑ Banco de España launched BELab in 2019 to provide the research community with secure access to high-quality microdata, both on-site and remotely. All datasets are primarily anonymized, and output control is applied to all results.
- ❑ In 2021–2022, BELab incorporated highly sensitive credit datasets, including Loans to legal persons (CIR) and Loans and Economic Indicators (CIR_CBI) microdata.
- ❑ A novel secondary anonymization procedure was developed to protect these datasets.
- ❑ In 2026, BELab received a substantially larger dataset covering Loans to individuals (CIR-PF), posing new privacy, scalability and utility challenges.
- ❑ We evaluate potential alternatives to the existing anonymization framework, including differential privacy and synthetic data.



SECONDARY ANONYMISATION OF CIR AND CIR-CBI DATASETS

- ❑ BELab developed a novel secondary anonymisation approach for the CIR dataset, as existing methods did not adequately address its privacy and utility requirements.
- ❑ The CIR and CIR-CBI datasets contain data on multiple loans per debtor, and yearly samples per loan.
- ❑ Available SDC tools (e.g. sdcMicro, μ -ARGUS) present several limitations for this use case:
 - They require data to contain **one single row per individual respondent**, which often is not the case.
 - They do not support **time-series data protection**. Data is protected for each year independently, which can allow the cancellation of suppressions.
 - They do not support a **joint analysis of categorical and numerical variables**. Debtors which are sensitive when combining numerical and categorical variables (e.g., sector and amounts) are not protected.
- ❑ A new method was therefore designed and implemented to support the joint anonymisation of categorical and numerical variables in sensitive time-series microdata:
 - [Joint secondary anonymisation of categorial and numerical variables in sensitive time series microdata](#). Eugenia Koblents and Alberto Lorenzo, Bank of Spain
- ❑ The new method has been successfully applied to CIR and CIR-CBI datasets since 2021.

SECONDARY ANONYMISATION OF CIR AND CIR-CBI DATASETS

- ❑ The implemented approach makes use of the open-source R package **sdcmicro** for risk assessment and microdata protection and Python for data pre and post processing:
 1. Identification of continuous and numerical debtor and loan **key variables** that can allow debtor re-id.
 2. **Global recoding** of selected key variables reducing the number of classes and disclosure risk.
 3. Creation of **full debtors' profiles** that incorporate information on all their loans throughout the full time series.
 4. **Debtor anonymisation**: local suppressions performed on debtor profiles with **sdcmicro** to guarantee k-anonymity.
 5. **Reintroduction of zeros** suppressed by **sdcmicro** into the debtor profiles.
 6. **Additional local suppressions based on nearest neighbors** to mitigate re-identification risks arising from the reintroduction of zero values.
 7. **Transfer of local suppression** patterns of debtors to the original **loans'** dataset.
- ❑ When **new yearly data** is incorporated to the dataset the full process needs to be repeated, requiring that each researcher only has access to one version of the dataset to avoid the cancellation of local suppressions.

SECONDARY ANONYMISATION OF CIR AND CIR-CBI DATASETS

1. Identification of categorical and numerical debtor and loan **key variables** that can allow debtor re-id in feasible disclosure scenarios. This is a challenging problem that needs to be addressed in collaboration with the data provider:

Debtor key variables	Loan key variables
Residence	Currency
Institutional sector	Personal guarantee
Economic activity	Investment region
Enterprise size	Drawn amount
Legal form	Undrawn amount

2. **Global recoding** of selected categorical debtor and loan key variables by grouping existing classes to significantly reduce disclosure risk. This process is agreed with the data provider and data users to guarantee high **data utility**.

Categorical variables	Original categories	Modified categories
Institutional sector (debtor)	16	3
Economic activity (debtor)	167	21
Currency (loan)	56	4
Personal guarantee (loan)	5	4
Investment region (loan)	55	18

SECONDARY ANONYMISATION OF CIR AND CIR-CBI DATASETS

3. **Debtors' profiles** are created, which contain information on all their **loans** extended in the whole time-series.

- **Categorical key variables describing loans** have been incorporated into the profile using **one-hot encoding** (an auxiliary binary variable has been created for each category of the original key variables).
- **Continuous key variables describing loans** are discretized according to the number of digits (loans with 1-6, 7, 8, 9, 10 and 11 digits) and are incorporated into the profile in the same way as categorical variables.
- As a result, **1.719.602 debtor profiles** (with one **single row per individual respondent**) are created containing **37 variables**: 5 debtor key variables and 32 one-hot encoded loan key variables.

4. **Debtor anonymisation**: local suppressions are performed on debtor profiles with **sdcmicro** to guarantee k-anonymity.

- 128.851 (7.49%) debtors don't satisfy k-anonymity (k=3). Local suppression takes 2 days.
- sdcmicro suppresses both 0s and 1s to enforce k-anonymity. In this setting, however, a 0 represents the absence of data for a given characteristic. Therefore, the suppression of 0s cannot be transferred to the original data.

Debtor ID	EUR	USD	GBP	Other currencies	Madrid	Catalonia	Andalusia
52364	1	0	1	0	1	1	0
76354	0	1	0	0	0	0	1
75345	1	1	0	0	0	1	0

Debtor ID	1-6 digits	7 digits	8 digits	9 digits	10 digits	11 digits or more
52364	1	0	1	1	0	0
76354	1	1	1	0	1	0
75345	1	1	0	0	0	0

SECONDARY ANONYMISATION OF CIR AND CIR-CBI DATASETS

5. Reintroduction of zeros suppressed by sdcMicro into the debtor profiles. This yields a low number of vulnerable debtors (6372, 0.37%), which don't satisfy k-anonymity.

Debtor ID	EUR	USD	GBP	Other currencies	Madrid	Catalonia	Andalusia
52364	1	0	1	0	1	1	0
76354	✗	1	0	0	0	0	1
75345	1	1	0	0	0	1	✗

Debtor ID	1-6 digits	7 digits	8 digits	9 digits	10 digits	11 digits or more
52364	1	✗	1	1	0	0
76354	1	1	1	✗	1	0
75345	1	1	0	0	0	0

6. Additional local suppressions are performed to protect those vulnerable debtors:

- Nearest neighbors of the 6372 debtors that don't satisfy k-anonymity are identified, and differing characteristics (1's) are suppressed to finally enforce k-anonymity, only suppressing 1 values (30 minutes)

Debtor ID	EUR	USD	GBP	Other currencies	Madrid	Catalonia	Andalusia
52364	1	0	✗	0	1	1	0
76354	✗	1	0	0	0	0	1
75345	1	1	0	0	0	1	✗

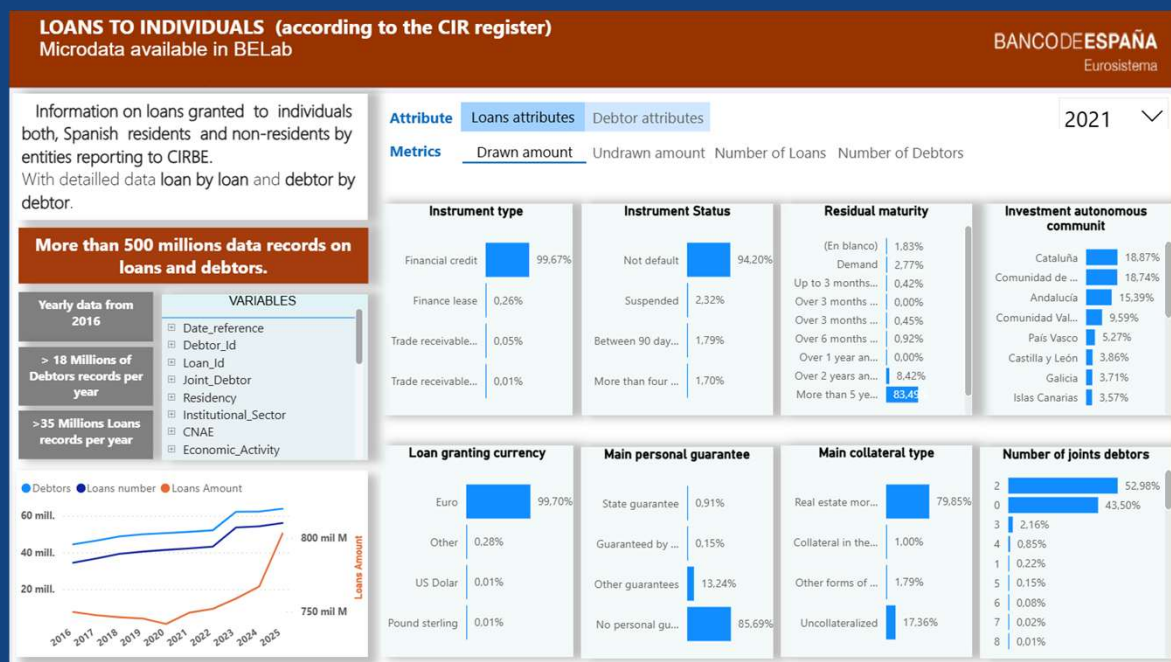
Debtor ID	1-6 digits	7 digits	8 digits	9 digits	10 digits	11 digits or more
52364	1	✗	1	1	0	0
76354	1	1	1	✗	1	0
75345	✗	1	0	0	0	0

7. Transfer of local suppressions from debtors' profiles to the original **loans'** dataset. Examples:

- For loans associated with debtor 52364 with currency GBP, the currency variable is suppressed.
- For loans associated with debtor 52364 with a 7-digit loan amount, the loan amount variable is suppressed.

SECONDARY ANONYMISATION OF THE CIR-PF DATASET

- ❑ In April 2026, BELab received a loans-to-natural-persons dataset approximately ten times larger than CIR and CIR-CBI.
- ❑ The existing anonymization framework did not scale efficiently to this new setting if applied to all debtors, motivating an assessment of alternative approaches such as differential privacy and synthetic data.



	CIR	CIR-PF
Dataset size (csv)	6 GB	106 GB
Dataset location	Local workstation	Spark Datalake
Total debtors	1.7 million	32.7 million
Total rows	43.7 million	531.8 million
Variables	21	22
Key variables	11	9

ALTERNATIVE METHODS: DIFFERENTIAL PRIVACY

- ❑ Differential Privacy (DP) has been considered as an alternative approach for protecting large-scale sensitive microdata.
- ❑ **Advantages of DP:**
 - ✓ Formal, mathematically provable privacy guarantees. Explicit and measurable privacy–utility trade-off.
 - ✓ Protection against attackers with external information.
 - ✓ Growing adoption by statistical agencies and technology companies.
- ❑ **Limitations for the BELab use case:**
 - ⚠ Primarily developed for **statistical query systems** and the release of **aggregate information**, rather than for direct researcher access to microdata
 - ⚠ Applying DP at the microdata level typically requires substantial perturbation, which may significantly reduce analytical utility.
- ❑ For BELab, the key challenge is protecting a relatively small set of high-risk observations while maintaining the analytical value of the underlying microdata. In this context, targeted anonymization methods were considered more practical and proportionate than Differential Privacy.

ALTERNATIVE METHODS: SYNTHETIC DATA

- ❑ The generation of a high-quality synthetic version of the dataset was considered as a potential solution for protecting large-scale sensitive microdata.
- ❑ However, this option was discarded based on the results of a previous Banco de España and European Commission pilot on synthetic financial data: [Utility and confidentiality assessment of synthetic company data, Eugenia Koblents](#)
- ❑ Main findings:
 - ✅ Strong privacy protection: no significant disclosure risks were identified.
 - ✅ Good preservation of basic statistical properties: univariate distributions and correlations
 - ⚠ Insufficient utility for research purposes:
 - Aggregate statistics were affected by the treatment of extreme values.
 - Important linear and non-linear relationships were not consistently preserved.
 - Utility varied substantially across research questions and analytical methods.
- ❑ Synthetic data provided strong confidentiality protection but did not preserve the analytical properties required for research-quality financial microdata. Therefore, it was not considered a suitable replacement for BELab datasets.

SELECTED APPROACH

- ❑ After evaluating alternative privacy-preserving approaches, it was decided to apply the previously developed anonymization method only to debtors with loans above €500,000 (this threshold can be adjusted as needed).
- ❑ The CIR-PF dataset is stored in a Spark-based data lake. Records corresponding to the most vulnerable debtors were extracted using Hive/Impala queries, anonymised separately, and then reintegrated into the original dataset.
- ❑ The resulting subset represented only a small fraction of the full dataset and could therefore be processed with low computational cost and acceptable execution times.
- ❑ Why this approach?
 - Natural persons are typically associated with less publicly available information than legal entities.
 - Protection efforts are focused on the debtors with the highest potential disclosure risk.
 - Good balance between privacy protection, analytical utility, scalability and implementation effort.

	CIR	CIR-PF
Dataset size (csv)	6 GB	106 GB
Dataset location	Local workstation	Spark Datalake
Total debtors	1.7 million	32.7 million
Processed debtors	1.7 million	0.13 million
Total rows	43.7 million	531.8 million
Processed rows	43.7 million	4 million
Variables	21	22
Key variables	11	9
Processing time	3 days	30 minutes
% suppressions	1.43	0.24

CONCLUSIONS AND FUTURE WORK

- ❑ BELab received a large-scale and highly sensitive credit microdata dataset covering loans to natural persons, requiring an additional layer of protection beyond standard access controls and output checking.
- ❑ Several Privacy-Enhancing Technologies (PETs) were evaluated, including Differential Privacy and Synthetic Data.
- ❑ While conceptually attractive, these approaches were not considered the most suitable solution for the BELab environment and use case.
- ❑ The previously developed anonymization procedure for CIR and CIR-CBI datasets was successfully adapted to the new dataset and applied to debtors with loan exposures above €500,000.
- ❑ By focusing on the most vulnerable observations, the approach concentrates protection efforts where disclosure risk is highest while preserving the analytical value of the data.
- ❑ The solution proved operationally feasible, computationally scalable and compatible with the large volume of records stored in the BELab infrastructure.
- ❑ Differential Privacy is being explored between BELab and Banca d'Italia for the generation of high dimensional confidential synthetic data and further applications (Daniele Piras and Alonso Sánchez)

Thank you for your attention!