BANCO DE ESPAÑA
Eurosistema

# JOINT SECONDARY ANONYMISATION OF CATEGORICAL AND NUMERICAL VARIABLES IN SENSITIVE TIME SERIES MICRODATA

## A novel approach for Statistical Disclosure Control of a microdata set published in BELab data laboratory

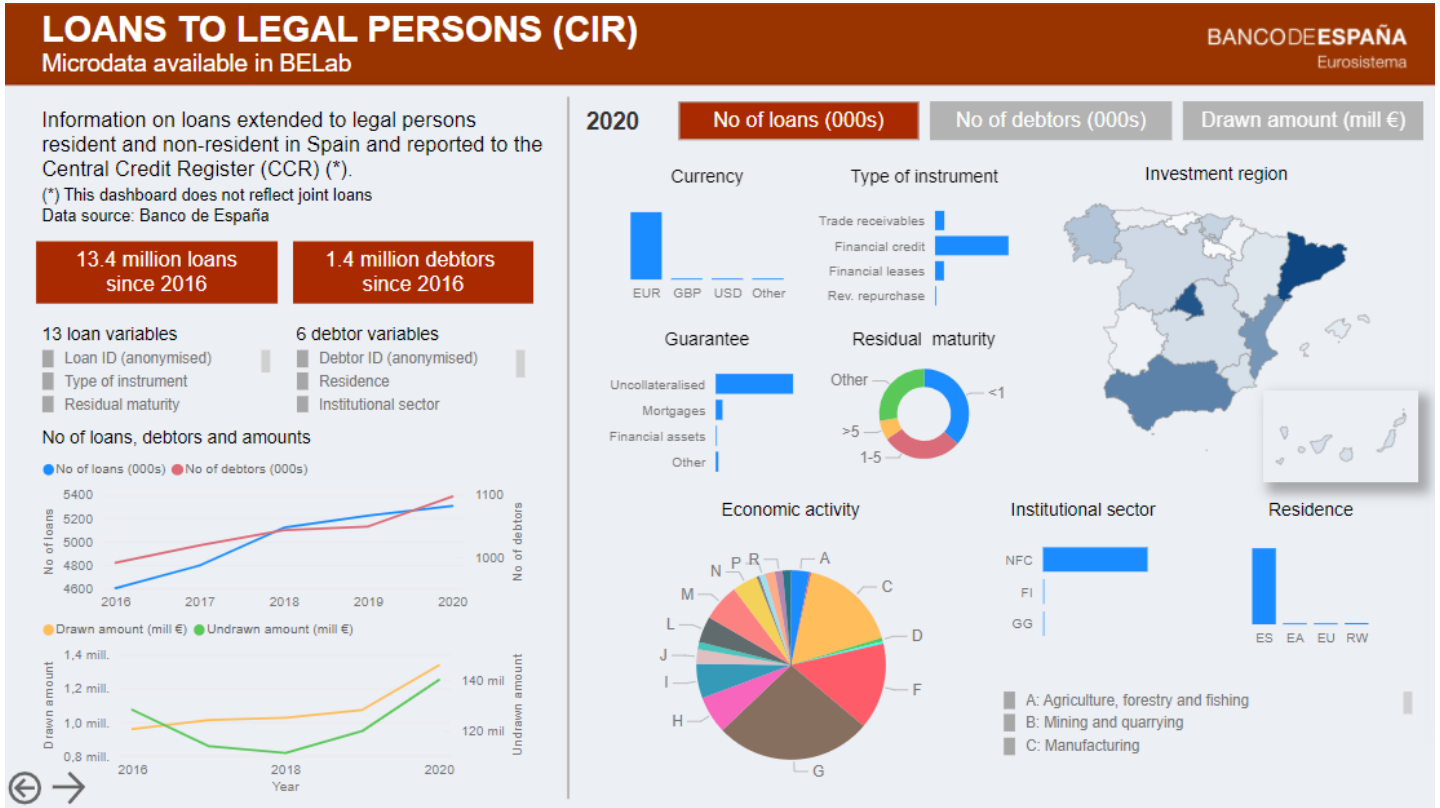
Eugenia Koblents
Alberto Lorenzo

ELEVENTH IFC CONFERENCE ON *"POST-PANDEMIC LANDSCAPE FOR CENTRAL BANK STATISTICS"*
BIS, BASEL, 25 AND 26 AUGUST 2022
INEXDA SESSION: MICRODATA DISCLOSURE CONTROL: A PRACTICAL PERSPECTIVE

25/08/2022

STATISTICS DEPARTMENT

- Banco de España launched **BELab** in July 2019 to provide access to the research community to high quality microdata via on-site and remote access. https://www.bde.es/bde/en/areas/analisis-economi/otros/que-es-belab/
- In October 2021 CIR (*Central de Información de Riesgos*) Department of Banco de España provided a very **sensitive dataset** to BELab containing information on **loans to legal entities** extended between 2016 and 2020.
- Primary and **secondary anonymisation** was required to protect debtor confidentiality.
- Categorical and numerical debtor and loan **key variables** have been identified.
- CIR dataset contains **multiple rows per debtor and loan** (joint loans).
- A novel SDC approach for secondary anonymisation has been designed and implemented which allows to **jointly analyse categorical and numerical key variables**.
- 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.



- Existing **SDC software tools** (sdcMicro, mu-argus) have some limitations which are relevant in this case:
  - They require data to contain **one single row per individual respondent**, which often is not the case.
  - They do not support a **joint analysis of categorical and numerical variables**.
  - Implemented anonymisation methods for **numerical key variables** (top/bottom coding, micro aggregation, noise addition, etc.) do not yield a good trade-off between disclosure risk and information loss for this problem.
  - They do not support **time-series data protection**.
- **A novel secondary anonymisation approach** has been designed and implemented which overcomes these difficulties:
  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. **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.

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 |

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).

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

- **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.

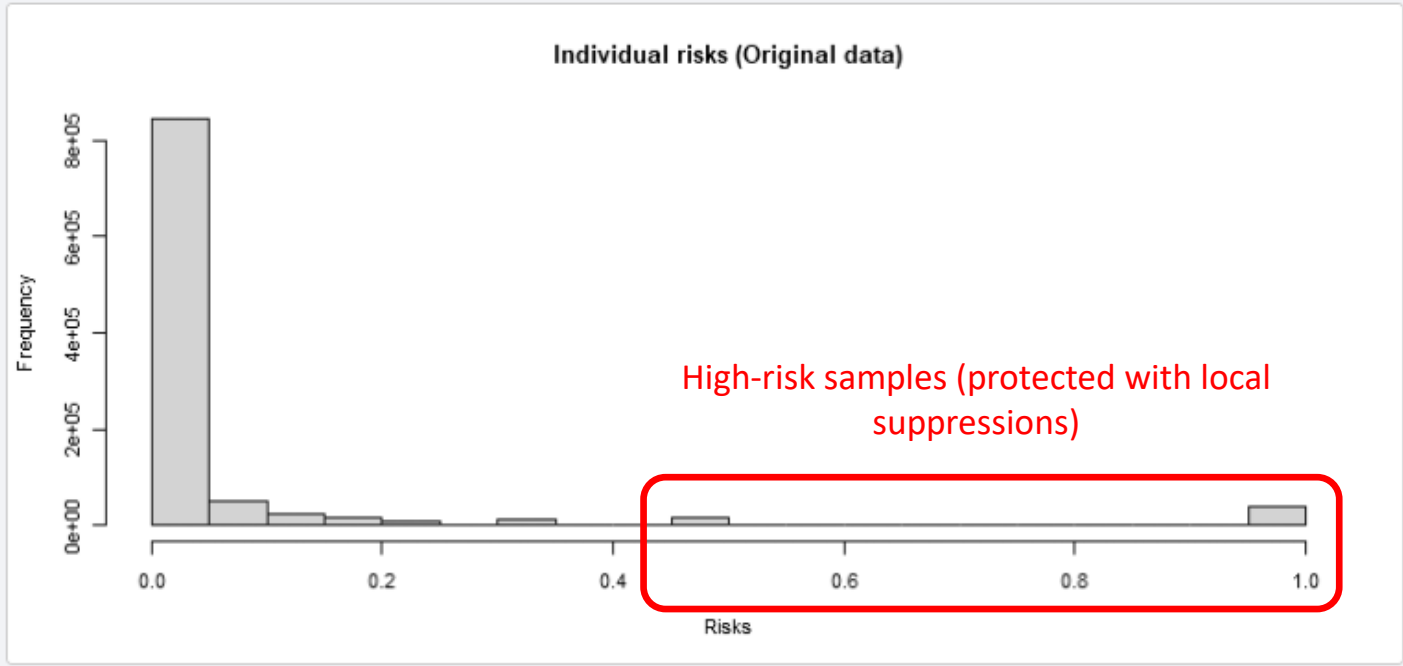
| Debtor ID | 1-6 digits | 7 digits | 8 digits | 9 digits | 10 digits | 11 digits or more |
|---|---|---|---|---|---|---|
| 52364 | 1 | 1 | 1 | 1 | 0 | 0 |
| 76354 | 1 | 1 | 1 | 1 | 1 | 0 |
| 75345 | 1 | 1 | 0 | 0 | 0 | 0 |

- As a result, **1.430.503 debtor profiles** (with one **single row per individual respondent**) have been created containing **37 variables**: 5 debtor key variables and 32 one-hot encoded loan key variables.

4. Debtor profiles contain **one single row per individual** and existing SDC software can be used to evaluate individual disclosure risks and to apply local suppressions to sensitive debtor profiles to guarantee **k-anonymity** (k=3).

| k-anonymity | Modified data | Original data |
|---|---|---|
| 2-anonymity | 0 (0.000%) | 46936 (3.281%) |
| 3-anonymity | 0 (0.000%) | 68450 (4.785%) |
| 5-anonymity | 14353 (1.003%) | 95733 (6.692%) |

Disclosure risk evaluation performed by **sdcMicro** before and after applying local suppressions to debtors' profiles.

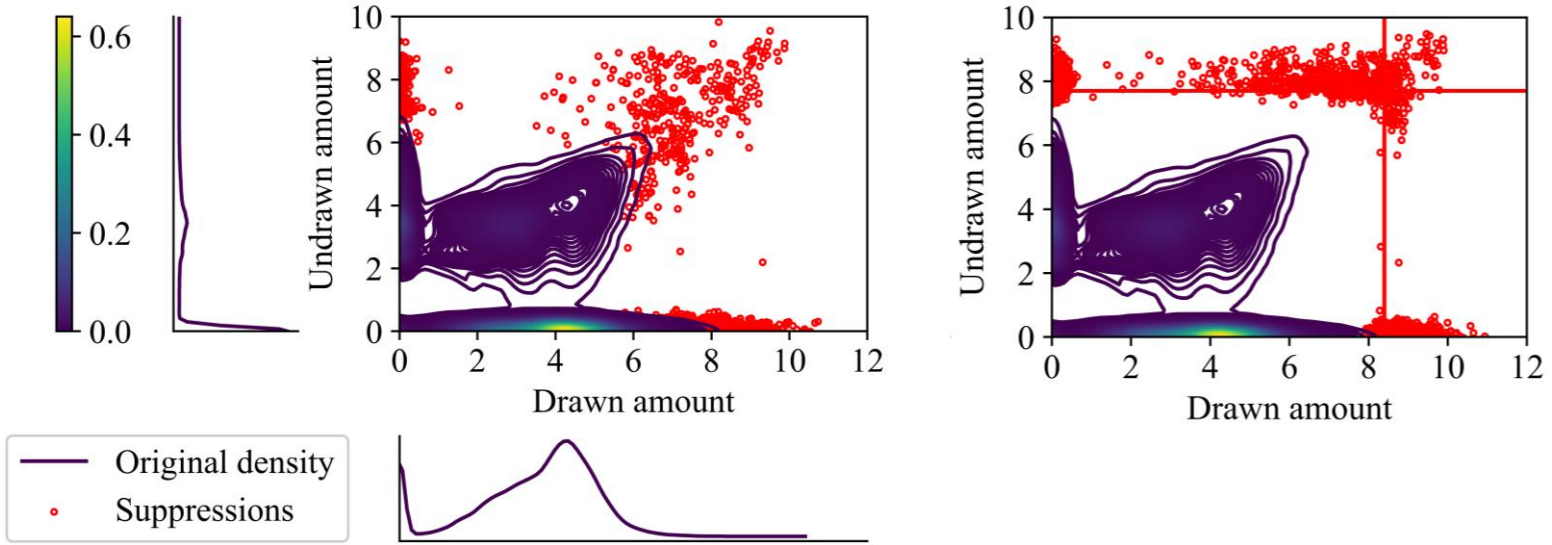



- **4.78%** of debtors are at risk of re-id in the original dataset, when combining categorical and numerical key variables.
- **0.15%** of debtors' information has been supressed (78.394 values of samples).
- As a result, all debtors in the anonymised dataset satisfy **k-anonymity** with k=3 (all individual risks are below 0.33).

5. Local suppressions obtained for debtors are **transferred** to the original loans dataset. **0.95%** of suppressions on average (mainly economical activity, legal form and company size). **0.01%** of numerical values suppressed.

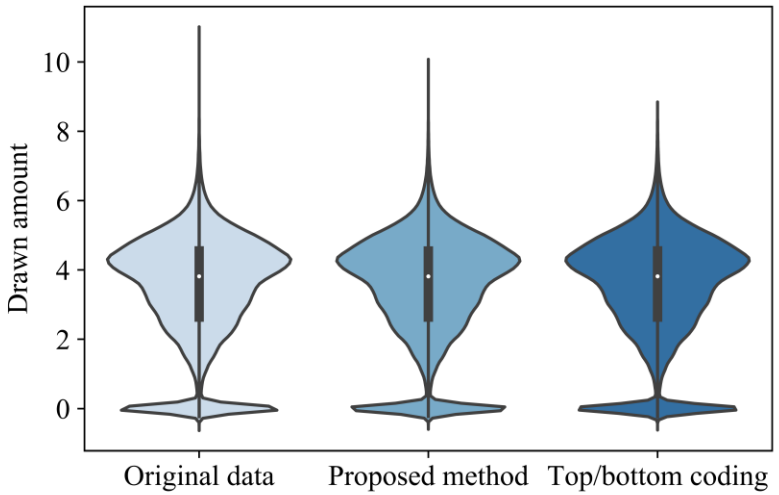
| Debtor and loan key variables | Number of suppressions | Percentage of suppressions |
|---|---|---|
| Legal form | 528,404 | 1.98 |
| Economic activity | 3,006,248 | 11.29 |
| Enterprise size | 447,127 | 1.68 |
| Drawn amount | 2,993 | 0.01 |
| Undrawn amount | 2,993 | 0.01 |
| TOTAL | **4,300,076** | **0.95** |

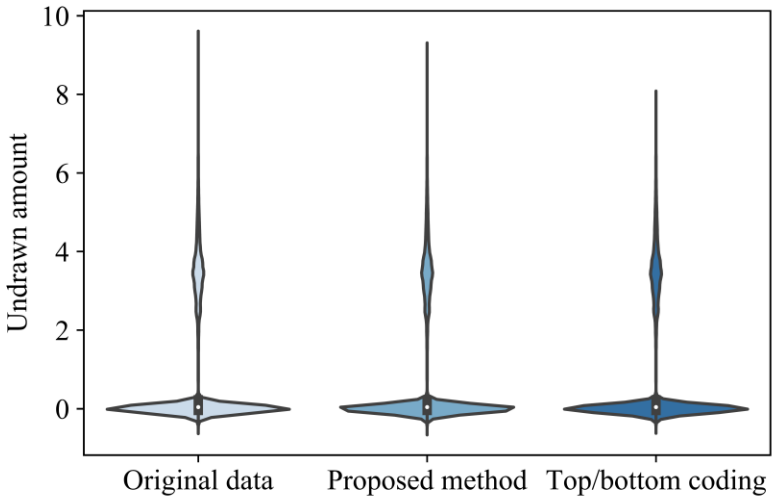
- Suppression pattern (red dots) obtained with the **proposed method** (left) and **top/bottom coding** (right).
- The proposed method protects samples that turn out to be sensitive when categorical and numerical variables are **jointly analysed** (largest loan per sector or region, etc.), while top/bottom coding consistently supresses **high-valued samples**.



- The proposed method affects the **data distribution and summary statistics** less than top/bottom coding. The obtained suppression pattern yields a **low information loss and disclosure risk**, since only sensitive samples are modified.

- Only the **tail of the distribution** is affected by anonymisation. The proposed method affects the tails less than top/bottom coding, because less outliers are suppressed.




- **Summary statistics** are significantly less affected using the proposed method than top/bottom coding.

| Summary statistics | Drawn amount | | Undrawn amount | |
|---|---|---|---|---|
| | Top/bottom coding | Proposed method | Top/bottom coding | Proposed method |
| Maximum | -99.6% | -83.7% | -98.4% | -53.0% |
| Mean | -24.0% | -12.9% | -36.1% | -11.7% |
| Standard deviation | -88.7% | -68.3% | -79.7% | -34.6% |
| Median | -0.03% | -0.04% | - | - |

# CONCLUSIONS

## And future lines of research

- A novel **secondary anonymisation** approach has been designed and implemented to protect the CIR dataset as a result of a close **collaboration** between BELab and the data provider (CIR Department of Banco de España).
- This procedure has a number of **benefits** over alternative procedures:
  - It allows to jointly analyse and protect **categorical and continuous variables**. Information on loans is incorporated into the debtors data, yielding a very complete **profile** for each debtor and allowing to use existing **SDC software**.
  - The joint anonymisation of categorical and numerical variables **minimizes disclosure risk and information loss** since only sensitive samples are affected. Only **0.95%** of suppressions (**0.01%** of numerical values).
  - The full **time-series** dataset is protected as a whole.
  - Once the described procedure has been designed, its **implementation** and use is relatively simple and its **computational cost** is low, in comparison with alternative methods, such as micro aggregation.
- Note that the full process needs to be repeated every time **new yearly data** becomes available. Researches cannot have access to more than one version of the dataset simultaneously, to avoid the possibility of cancelling local suppressions.
- **Future research lines**:
  - Address the possibility of **modelling uncertainty** on the information available to intruders.
  - Analyse links between **microdata protection and anomaly detection**, since both topics are closely related.

BANCO DE ESPAÑA
Eurosistema

# Thank you for your attention!

STATISTICS DEPARTMENT