IFC Satellite Seminar on "Granular data: new horizons and challenges for central banks"

Utility and confidentiality assessment of synthetic company data[1]

Eugenia Koblents Lapteva,
Bank of Spain

[1] This contribution was prepared for the IFC Satellite Seminar held at the ISI 64th World Statistics Congress, co-organised with the Bank of Canada in Ottawa, Canada, on 15 July 2023. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Canada, the BIS, the IFC or the other central banks and institutions represented at the event.

BANCO DE ESPAÑA
Eurosistema

UTILITY AND CONFIDENTIALITY ASSESSMENT OF SYNTHETIC FINANCIAL DATA

Synthetic data pilot in collaboration with the European Commission

Eugenia Koblents
Banco de España, Statistics Department, BELab data laboratory

BOC-IFC SATELLITE SEMINAR ON "GRANULAR DATA: NEW HORIZONS AND CHALLENGES FOR CENTRAL BANKS"

Bank of Canada, July 15, 2023

STATISTICS DEPARTMENT

INDEX

INTRODUCTION

Synthetic data pilot with the European Commission

- This presentation outlines Banco de España's participation in the **synthetic data pilot** launched by the European Commission in the summer of 2022, to support the creation of the EU Digital Finance Platform's **Data Hub**.

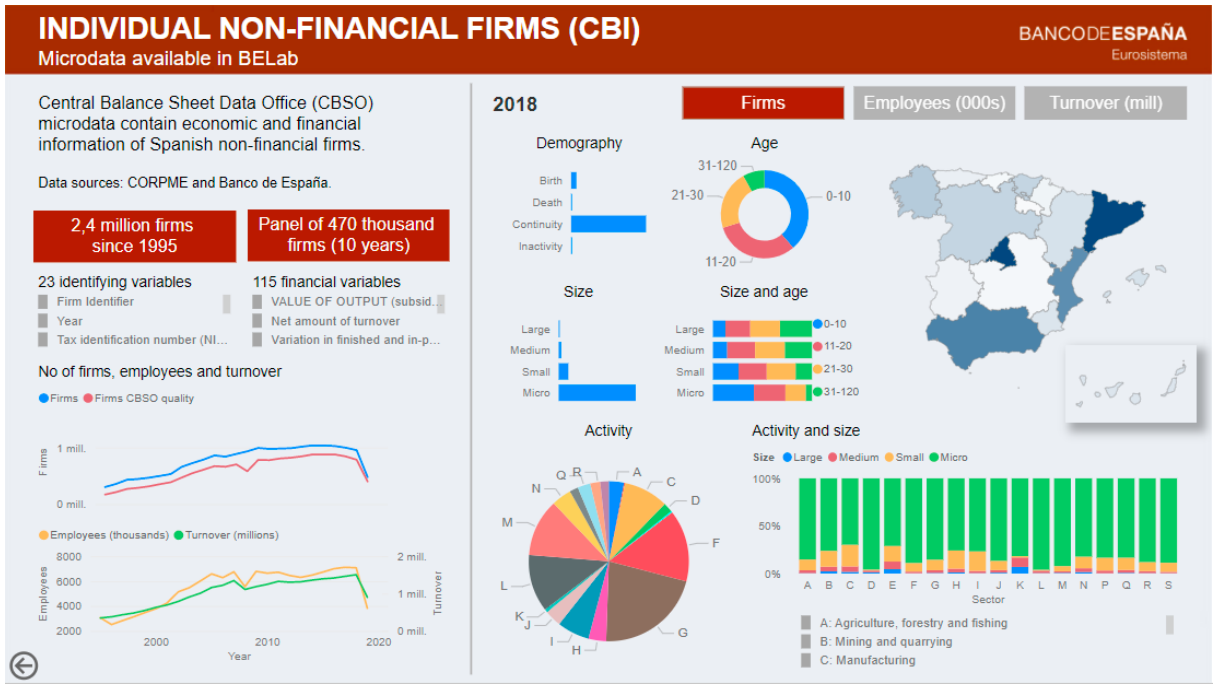
- The Data Hub aims to provide participant FinTech companies with access to non-public datasets to test their solutions. However, the data is often **highly confidential** and cannot be shared with third parties due to legal restrictions.

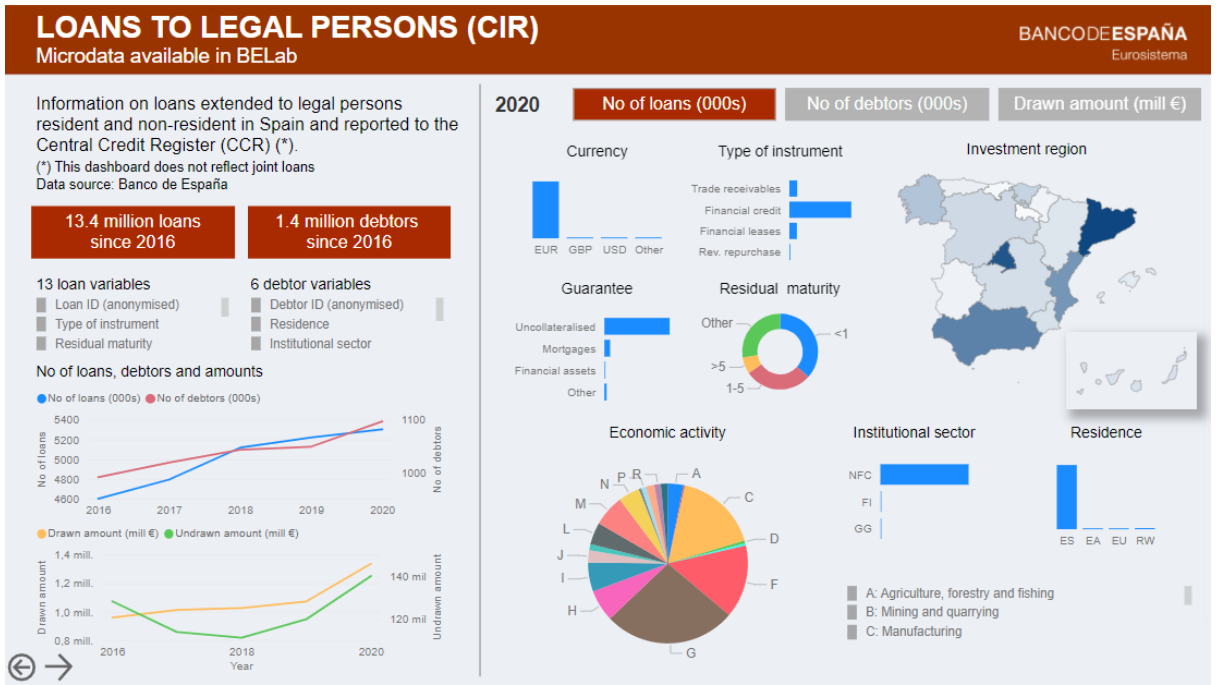
- **Synthetic data** generation was proposed as an alternative solution to enable **sharing of sensitive information without compromising privacy**. Synthetic data preserves the characteristics of the original data while ensuring confidentiality.

- Therefore, the pilot aimed to assess the **accuracy and privacy** of synthetic data and determine its suitability for different applications such as research, statistical analysis, IT development testing, and auditing.

- **Pilot participants**:
  - **European Commission**: Pilot leader and coordinator, contracting authority.
  - **AI company:** Synthetic data software and training provider.
  - **Banco de España**: Data provider, synthetic data assessment.

- **BELab** data laboratory and datasets used: https://www.bde.es/bde/en/areas/analisis-economi/otros/que-es-belab/
- **CBI** contains economic and financial variables built from the **annual accounts** reported by Spanish non-financial firms. Almost one million samples per year, 70 selected categorical and numerical variables.
- **CIR** contains sensitive information on **loans to legal entities** resident and non-resident in Spain. 5 million samples per year, 19 categorical and numerical variables. Multiple time series per debtor and loans with multiple debtors (joint loans).





INTRODUCTION

Data description and performed synthetization

- The software tool used implements the **full pipeline of fully synthetic data generation:** data upload, model training parameters configuration, execution supervision, and results analysis. The tool integrates a web-based user interface.

- According to the utility and privacy metrics computed by the software, **CBI and CIR datasets** have been successfully synthesised **independently and jointly**:

Synthesised datasets		Number of variables	Number of samples	Accuracy	Privacy tests
CBI	Static	8 numeric, 3 categorical	1,238,393	98%	Passed
	Dynamic	55 numeric, 1 categorical	4,067,062	95%	Passed
CIR (anon)	Static	5 categorical	699,480 (sample)	99%	Passed
	Dynamic	5 numeric, 7 categorical	10,434,349	98%	Passed
CBI-CIR	Static	8 numeric, 8 categorical	100,000 (sample)	99,4%	Passed
	Dynamic	59 numeric, 8 categorical	560,056	97,5%	Passed

- **Additional tests** have been conducted at Banco de España to complement the evaluation analysis.

- CBI and CIR datasets have been used for **utility and privacy assessment**, respectively.

INDEX

UTILITY ASSESSMENT

Utility metrics for fully synthetic data

- Evaluating the utility of the generated data is a pivotal step in any synthetic data project, to identify the **most suitable synthesis strategy** and to decide whether the data are of **sufficient quality** to be released.

- **Utility metrics for fully synthetic data can be broadly divided into three categories:**

  - **Global utility metrics** assess the utility by directly comparing the original and the synthetic data (Kullback-Leibler divergence, propensity score matching, etc.), making no assumptions on the type of analyses to be conducted.

    Utility is measured on a **very aggregated level**, so good results do not guarantee high utility for specific analyses.

  - **Outcome-specific utility metrics** measure the utility for a specific analysis (means or regression coefficients for original and synthetic data, confidence interval overlap, etc.).

  - **Fit-for-purpose measures** provide a first impression of the quality of the synthetic data (marginal and bivariate distributions, consistency checks, goodness-of-fit measures, etc.).

[1] Drechsler, J., & Haensch, A. C. (2023). 30 Years of Synthetic Data. *arXiv preprint arXiv:2304.02107*.

[2] Drechsler, J. (2022, September). Challenges in Measuring Utility for Fully Synthetic Data. PSD 2022, Paris.

[3] Synthetic Data for Official Statistics, UNECE, 2022: https://unece.org/sites/default/files/2022-11/ECECESSTAT20226.pdf

UTILITY ASSESSMENT

Fit-for-purpose and outcome-specific utility metrics

- The synthetic data software implements several **graphical fit-for-purpose measures** (comparison of univariate and bivariate distributions, correlation matrices and pairwise accuracy matrix), yielding very good results in this experiment.

- Furthermore, the team at Banco de España has conducted additional **outcome-specific utility analysis** to evaluate the quality of the generated CBI synthetic data, with a focus on Banco de España's use cases.

- Two critical factors determine the utility of the synthetic **CBI dataset** for common use cases:

  - Accurate replication of the original **statistical distributions**: To validate this, various aggregated **summary statistics** obtained from real and synthetic data have been compared.

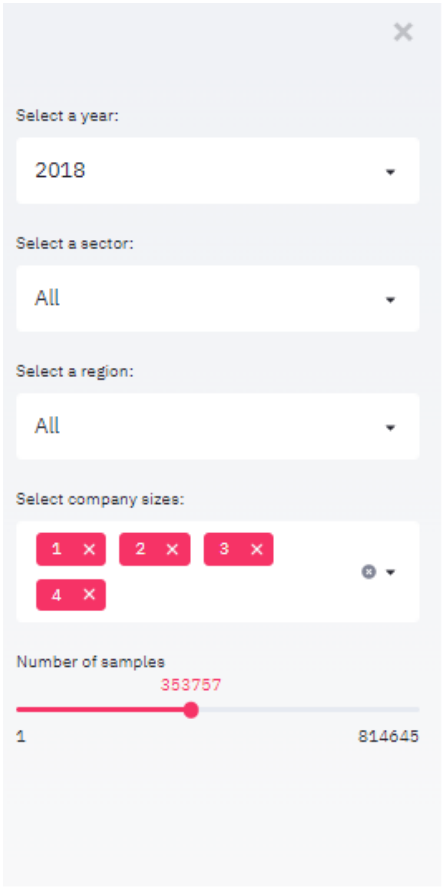
  - Preservation of **arithmetic relationships** among variables describing balance relationships: This aspect has been evaluated by comparing the **unbalance** between real and synthetic variables, respectively.

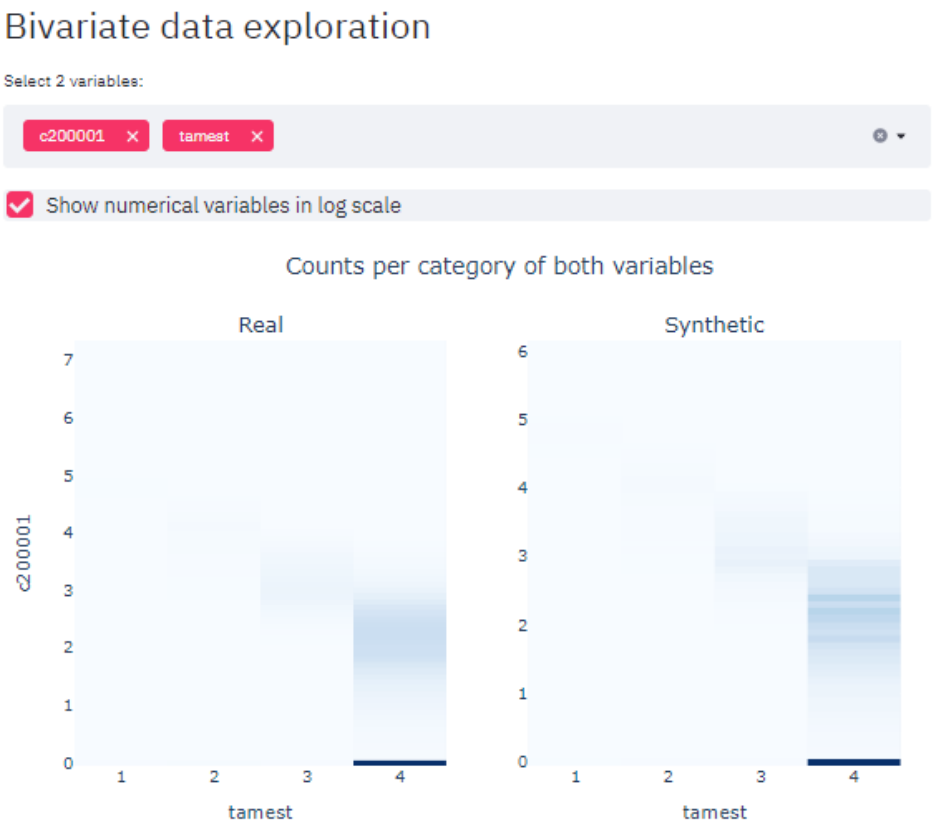
- Additional **comprehensive utility assessments** have been conducted by the **Suptech** and **Financial Innovation** teams at Banco de España, involving the comparison of machine learning models trained on real and synthetic data, descriptive time series analysis, and adversarial validation [4], reaching similar conclusions to the ones presented here.

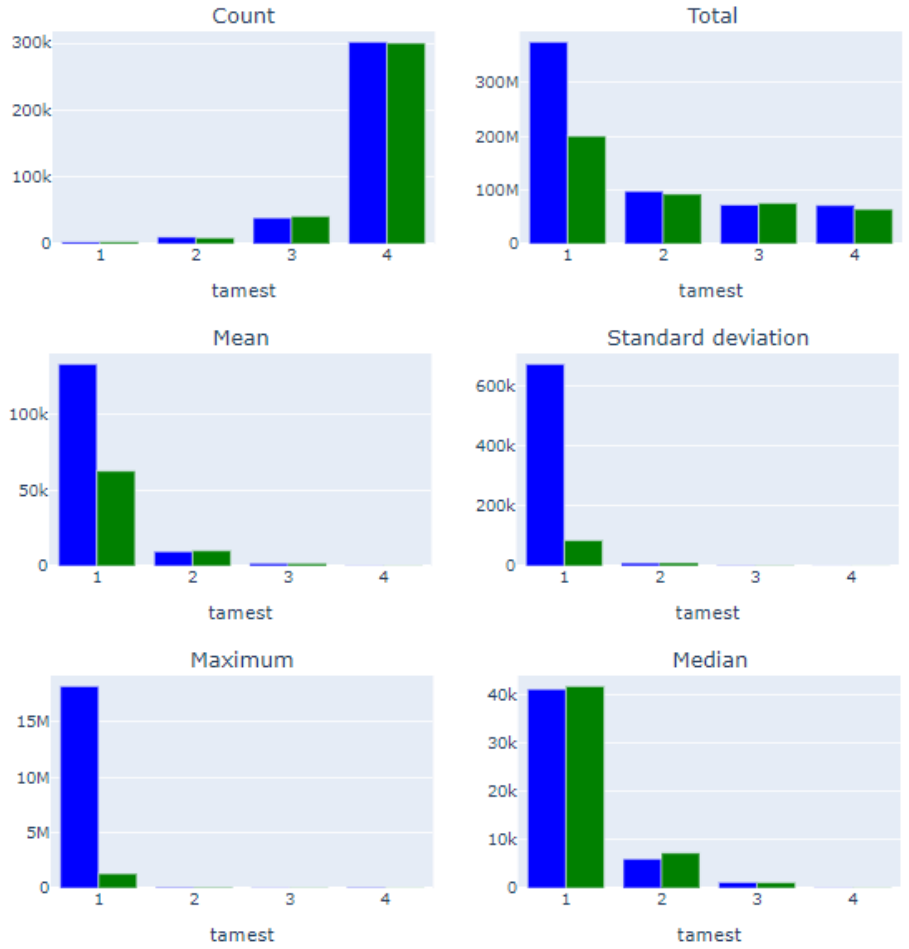
[4] Alonso, A., & Carbó, J. M. (2022). Accuracy of Explanations of Machine Learning Models for Credit Decision.

- **Global aggregates computed for BELab dashboards**: deviations in total amounts are observed, specially for large companies. This is due to **outlier suppression,** which is performed in order to avoid membership inference attacks.

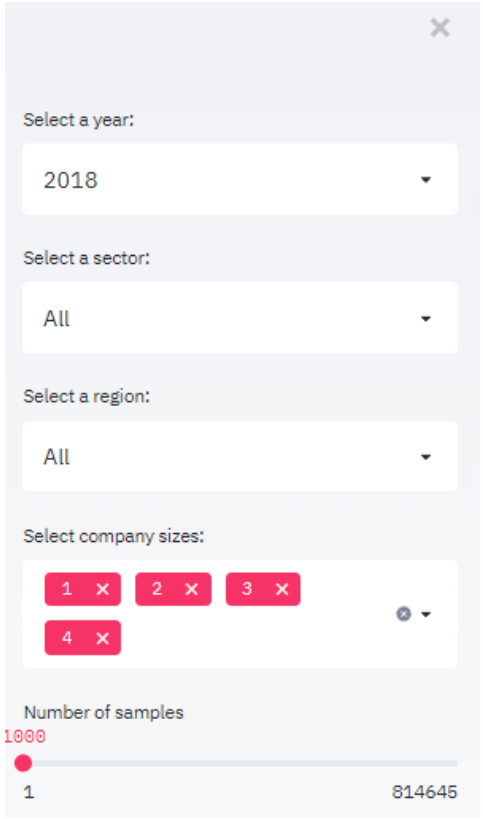
- **Analysis of summary statistics for sales per company size**. The number of companies (count) and the median are well reproduced but deviations are observed in the total, mean, maximum and std values.

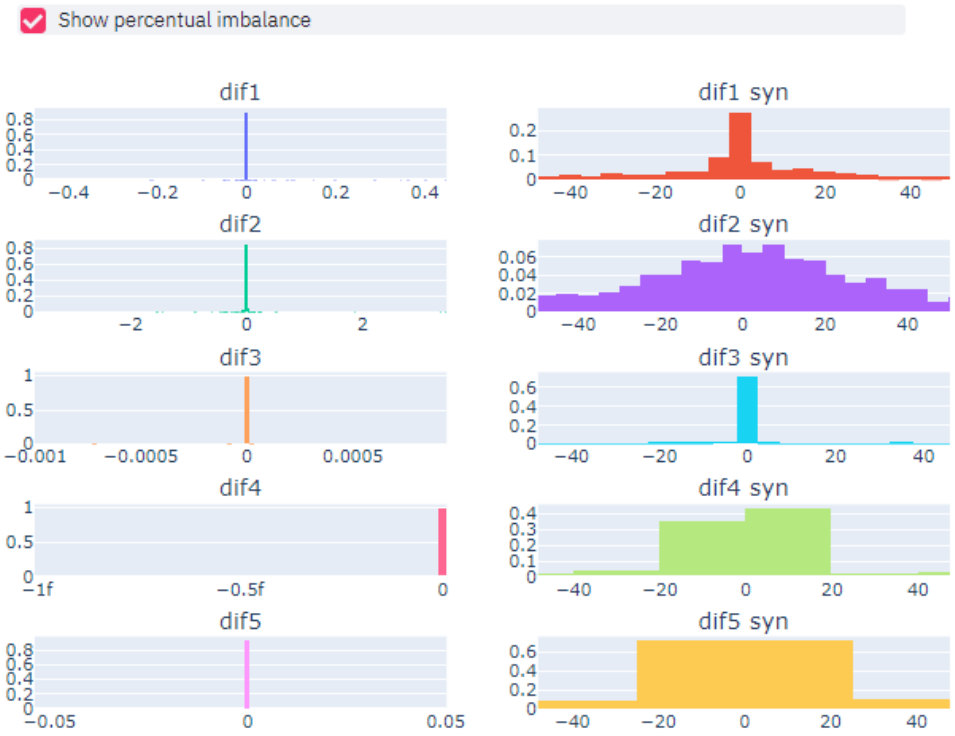






- **Analysis of linear relationships:** c200135 (total assets) = c200115 (non-current assets) + c200134 (current assets)
- In the real dataset most deviations are below 0.2% while in the synthetic dataset many samples present deviations above a 20%. Only **6% of samples** present an imbalance below 10% for these five linear relationships (many more exist).





Analysis of linear relationships in the data

Total imbalance 1 (c200115 = c200089 + c200098 + c200103): 835 vs 560346601

Total imbalance 3 (c200145 = c200142 + c200143 + c200144): 72 vs 423889005

Total imbalance 4 (c200158 = c200146 + c200151): 1 vs 239324366

Total imbalance 5 (c200180 = c200159 + c209166 + c209179): 716 vs 266960469

Total imbalance 6 (c200135 = c200115 + c200134): 1668 vs 748815588

Discard null values

Percentage of companies with an imbalance below a thresold

Maximum percentage of imbalance

10

Balance 1: 99 vs 54

Balance 2: 99 vs 26

Balance 3: 99 vs 80

Balance 4: 100 vs 80

Balance 5: 99 vs 59

Balance 6: 99 vs 74

Combination of all: 99 vs 6

INDEX

CONFIDENTIALITY ASSESSMENT

Disclosure risk in fully synthetic data

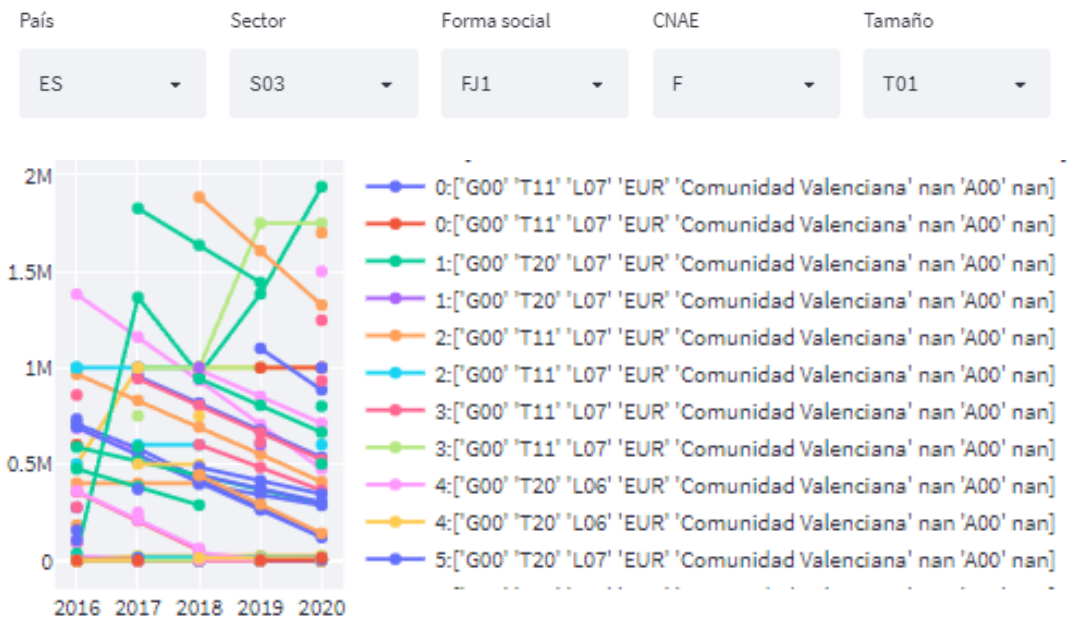
- No record in a fully synthetic data file corresponds to a real individual. Therefore, measuring the **risk of re-identification** is not applicable to fully synthetic data [1].
- However, National Statistical Authorities (NSA) often resort to disclosure risk metrics to avoid **reputational risk** [3]. Respondents might be concerned if they come across a synthetic record that exactly matches their own, especially if their combination of attributes is unique.
- Common disclosure risk measures for fully synthetic data include evaluating the **distance** between synthetic and real data records and the proportion of **unique matches**. In order to minimize reputation loss, these unique matches are often removed from the synthetic dataset.
- The synthetic data software employs various **distance-based metrics** to identify synthetic samples that closely resemble real ones. The synthesized datasets have **successfully passed** all of these tests.
- Furthermore, at Banco de España, **unique matches** in the CIR dataset have been identified to ensure that no real individual can be re-identified from the synthetic dataset to help assess and minimize **reputational risk**.

CONFIDENTIALITY ASSESSMENT

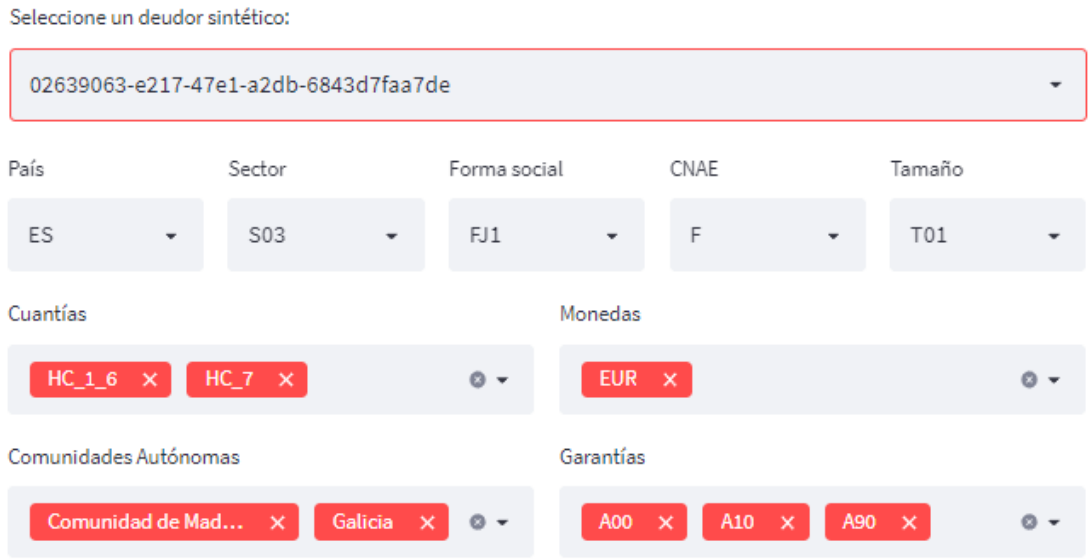
Unique matches between real and synthetic data samples

- The CIR dataset consists of **static debtor variables** (company size, sector of activity, etc) and a variable number of **time-series** describing loans, with variable length and attributes (amounts, currency, region of investment, etc).
- To facilitate distance computation, fixed-length **debtor profiles** were created and compared. Summarized time-series information was incorporated using one-hot encoding.

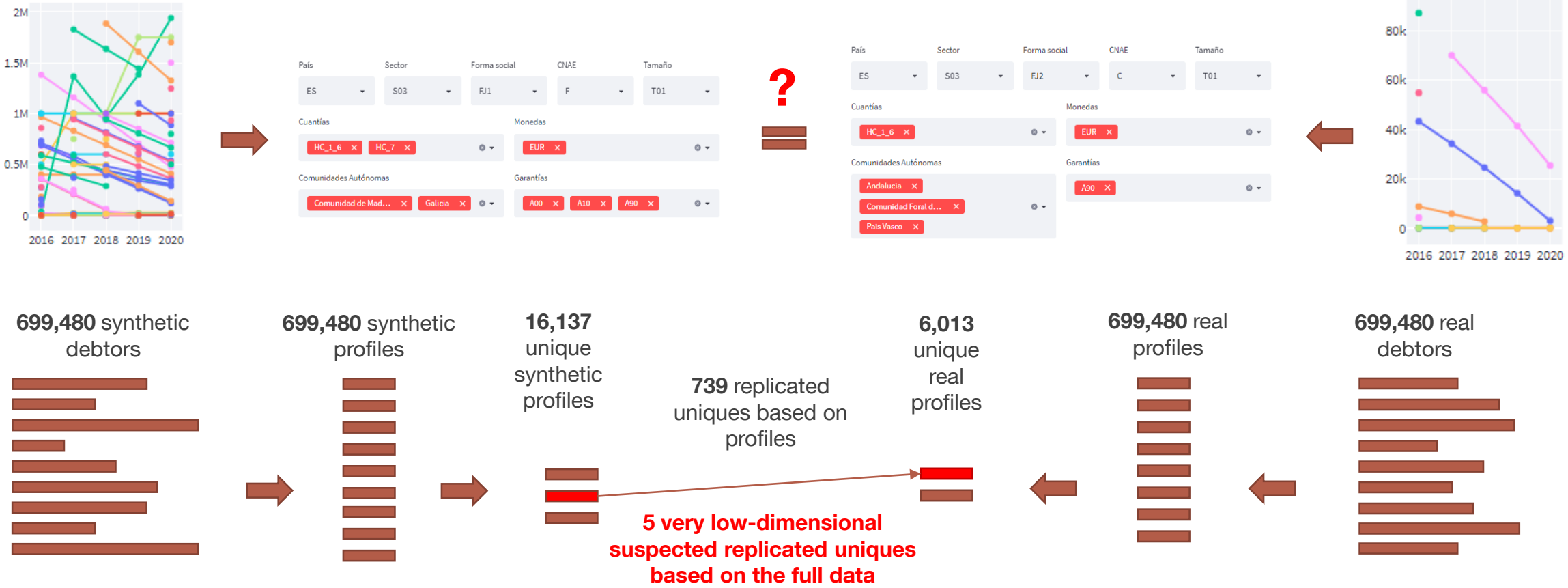
Full debtor's information



Summarized debtor's profile



- Fixed-length low-dimensional **debtor profiles** were created and compared.



- As the **dimension** of time series increases, the probability of exact matches decreases exponentially.
- Unique matches do not necessarily imply disclosure risk, but they have been supressed to **minimize reputational risk**.

INDEX

- A **pilot on synthetic data generation and evaluation** has been conducted at Banco de España in collaboration with the EU Commission and a private synthetic data generation software provider.
- The **main outcomes** of the pilot are the following:
  - The software allows to accurately reproduce **univariate and bivariate distributions** as well as the **correlations** present in the real data and guarantees the **privacy** of individual samples.
  - However, the provided **fit-for-purpose utility metrics** do not guarantee high utility for any analysis, which needs to be evaluated for each dataset and use case. Specific observations:
    - Outlier suppression heavily affects **summary statistics**. Datasets consisting of SMEs only and variables describing ratios instead of absolute values appear to be less sensitive to outlier suppression.
    - **Linear and non-linear relationships** among variables are often not preserved.
    - **Subsequent analyses** of the generated synthetic data have yielded similar conclusions regarding data utility [4].
  - **Merging synthetic datasets** with different, correlated variables is not possible with the technology used.
  - **No significant privacy issues** have been identified. Only a very low number of low-dimensional unique matches have been identified, which should be supressed to minimize reputational risk.

BANCO DE ESPAÑA
Eurosistema

Thank you for your attention!

STATISTICS DEPARTMENT