ISSN 1831-9424

Synthetic data in the Data Hub of the Digital Finance Platform

Di Girolamo, F., Hledik, J., Pagano, A.

2024

Joint
Research
Centre

EUR 31919 EN

EU Science Hub
https://joint-research-centre.ec.europa.eu

JRC137249

EUR 31919 EN

PDF ISBN 978-92-68-15291-1 ISSN 1831-9424 doi:10.2760/83055 KJ-NA-31-919-EN-N

Luxembourg: Publications Office of the European Union, 2024

How to cite this report: European Commission, Joint Research Centre, Di Girolamo, F., Hledik, J. and Pagano, A., *Synthetic data in the Data Hub of the Digital Finance Platform*, Publications Office of the European Union, Luxembourg, 2024, https://data.europa.eu/doi/10.2760/83055, JRC137249.

Contents

Abstract

The Digital Finance Strategy focuses on fostering a competitive and innovative European financial sector by leveraging emerging trends and technologies. As part of this effort, the EU Digital Finance Platform initiative aims to provide practical tools to support the scaling up of innovative financial firms across the EU. Additionally, the establishment of a Data Hub within the platform will facilitate data exchange between national supervisors and financial firms. To ensure compliance with confidentiality requirements, the European Commission has decided to build the Data Hub using synthetic data. The report describes the datasets used to test the methodology chosen for the synthetization. In addition, the report compares the main statistical properties of the original and new database, and summarizes the tests performed to draw conclusions on potential confidentiality and privacy issues. By testing and validating the data synthesis software, the JRC and DG FISMA are working to ensure that the new dataset will be a valuable resource for firms and researchers, while also respecting confidentiality issues.

1. Introduction

The EU digital finance agenda aligns with the broader Commission policy on digital transition and aims to preserve a level playing field across the financial sector and address challenges and risks associated with digital transformation. While it presents significant opportunities for the financial sector, it also requires supervisors to keep pace with the rapid changes to ensure financial stability, consumer protection, and market integrity.

In September 2020, the European Commission adopted a **digital finance package**, which includes the **Digital Finance Strategy.** The goal is to foster a more competitive and innovative European financial sector, by leveraging emerging trends and technologies to enhance Europe's competitiveness. By making rules more digital-friendly and safe for consumers, the strategy sets out four main priorities: removing fragmentation in the Digital Single Market, adapting the EU regulatory framework to facilitate digital innovation, promoting a data-driven finance, and addressing the challenges and risks associated with digital transformation.

As part of this effort to promote innovation in finance and establish a single market for digital financial services, **the EU Digital Finance Platform initiative** was already announced in the Strategy. The platform, set up by the Digital Finance Unit within the Directorate-General for Financial Stability, Financial Services and Capital Markets Union (DG FISMA), serves as a collaborative space that develops closer relationship between innovative financial institutions and national supervisors. The goal is to provide practical tools to support the scaling up of innovative financial firms across the EU, addressing the challenges faced by the fintech community when expanding their offerings across Europe. The platform provides a mapping to showcase Europe's fintech ecosystem. Additionally, the European Forum for Innovation Facilitators provides cross-border services for supervisors to share experiences, technical expertise, and foster exchange and mutual learning.

In line with the Commission's priorities of fostering the development of trustworthy data-sharing systems and enabling data-driven innovations, the strategy envisages the establishment of a **Data Hub** within the Digital Finance Platform. In addition to complementing national innovation hubs and sandboxes, the Data Hub will serve as a place for data exchange between national supervisors and financial firms. This initiative aims to match data needs within the financial sector with datasets held by national supervisors to allow participating companies to test innovative solutions and train AI/ML models. It represents a crucial step for a robust and innovative digital finance landscape, where collaboration and technological advancement are the cornerstones of progress. On one hand, the Data Hub will provide participating firms, academics and researchers, with access to specific sets of supervisory data for testing new solutions and training AI/ML models in collaboration with supervisors. On the other hand, by supporting this project, national supervisors will not only encourage innovation but also gain insights into the technologies used by innovative firms.

To ensure compliance with confidentiality requirements, the European Commission has decided to build the Data Hub using synthetic data. Synthetic data, which is artificial data generated from

original data and trained to reproduce the characteristics and structure of the original data, offers a way for national supervisors to participate in the project without having to make the real data they hold accessible to any third party.

The national supervisors will use a software provided by a private firm to generate synthetic data from the original data they hold. The new dataset will offer the necessary level of anonymization, while preserving the characteristics of the original data that make it relevant for testing purposes. At no point will any real data leave the premises of the respective supervisors, nor will any external user gain access to it. The European Commission has acquired the services related to the creation of synthetic data through a tender procedure, ensuring that all participating authorities use the same program and the resulting synthetic datasets have the same format.

The responsibility for building the Data Hub falls under the Digital Finance Unit of DG FISMA, while the JRC is working to test the data synthesis software. The purpose of this testing is to ensure that the new datasets maintain the properties of the original dataset while also protecting privacy and confidentiality. The report describes the datasets used to test the methodology and the steps taken for the synthesis. In addition, the report compares the main statistical properties of the original and new database, and summarize the tests performed to draw conclusions on potential confidentiality and privacy issues. By testing and validating the data synthesis software, the JRC and DG FISMA are working to ensure that the new dataset will be a valuable resource for firms and researchers, while also respecting confidentiality issues.

2. Data description

The analysis in this report is based on three individual datasets. The first dataset contains individual banks' balance sheet data, which has been sourced from Moody's Analytic BankFocus.[1] The second dataset contains retail bank account transactions from an anonymous European commercial bank, while the third dataset describes interbank lending between Austrian commercial banks. We mostly focus our attention on the first dataset, with additional remarks obtained from the other two datasets.

Dataset 1: Banks' balance sheets

The final data set covers 250 EU banks located in EU 14,[2] including commercial, saving and cooperative banks. The data refers to the end of 2022 and includes a comprehensive range of balance sheet items, summarized in Table 1. These variables include levels (such as total assets, net loans, risk weighted assets), as well as ratios (such as Tier 1 ratio). We have excluded financial institutions from the sample if they had missing values in any of the variable of interest listed below. The decision was made to ensure the integrity of our analysis. Missing data could in fact impact the accuracy of our findings after the synthesis potentially leading to biased results.

[1] https://www.moodysanalytics.com/product-list/bankfocus

[2] For the other EU countries, we do not have banks without missing values.

Table 1: Variables of interest in the banks' balance sheets database.

Main	Secondary	Ratios
Total assets	Net loans	Tier1 over risk weighted assets
Total liabilities	Loan loss reserves	Equity over total assets
Equity	Derivatives	Risk weighted assets over total assets
Tier 1 capital	Securities and investments	Gross loans over total assets
Risk weighted assets	Other earning assets	Interbank assets over total assets
Gross loans	Total earning assets	
Interbank assets	Fixed assets	
	Total other earning assets	

Dataset 2: Retail bank account transactions

This dataset contains 105 316 observations of individual retail transactions from an anonymous European commercial bank. It spans the period of one calendar year with information from 100 randomly picked retail bank accounts. For each transaction, the available variables are the relevant date, transaction amount in EUR and a binary variable specifying whether it's a credit or debit transaction.

Dataset 3: Interbank lending

This dataset contains quarterly observations of interbank loans between 800 Austrian commercial banks. The period spans 4 years between 2008 and 2012. Due to privacy reasons, the banks are anonymized and referred to by numbers between 1 and 800.

There are four variables available within this dataset: the period (quarter), the number of the lender bank, the number of the borrower bank and a binary variable expressing whether there is a lending relationship between the two banks in this specific period. Further details about the dataset can be found in Puhr et al. (2012)[3] and Hledik and Rastelli (2023)[4].

3. The synthesis process

The data synthesis procedure has been outsourced to Synthesized, an external counterparty chosen as a result of a public tender within the Data Hub project. Synthesized is providing a comprehensive data synthesis software with their Synthesized SDK. The JRC was provided with

[3] Puhr, C., R. Seliger, and M. Sigmund, 2012, Contagiousness and vulnerability in the austrian interbank market, Oesterreichische Nationalbank Financial Stability Report 24

[4] Hledik, J. and Rastelli, R., A dynamic network model to measure exposure concentration in the Austrian interbank market, STATISTICAL METHODS AND APPLICATIONS, ISSN 1618-2510, 2023, JRC134154.

a temporary license to the Synthesized's SDK in order to conduct the analyses and tests summarized in this document.

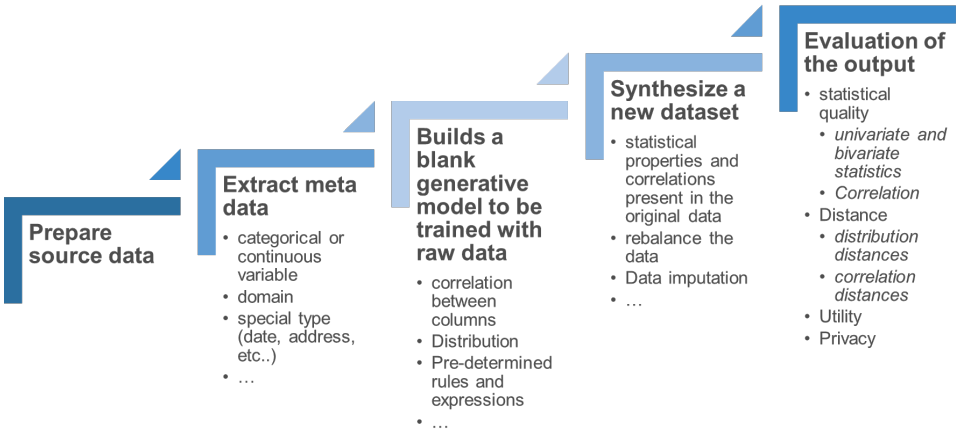
The relevant SDK takes form of a commercial Python package. For our purposes, it has been tested on a Ubuntu Linux machine with 24 CPU cores and 64 GB or RAM. Usage of a GPU during the synthesis mechanism is possible but has not been explored given the already short timeframe of the process. Notwithstanding, none of the datasets which we have explored took longer than a few minutes to synthesize while using the CPU. Installation of the synthesis SDK package is very straightforward. It can be done with 2 simple commands in a linux terminal and a third command with a license key needs to be entered to register the product afterwards. The actual synthesis is also very simple (see Figure 1). First, the original dataset's metadata is extracted. Second, it is used to instantiate the model. Third, the model is trained on the original dataset and fourth, new data is synthesized. The SDK allows user to select specific dependencies within the dataset. For instance, it is possible to force the synthesized data to follow desired algebraic equalities across specific variables, or to specify bounds on any of them. If the original dataset exhibits such deterministic relationships among its variables, it is very easy to replicate in the synthesized data. [5]

In terms of methods, the Synthesized SDK uses Tensorflow, which is a popular open-source deep learning package developed by Google. From an outside perspective, this is a black box which likely contains several different machine learning algorithms to create the desired synthetic dataset. Deep learning algorithms are known for their opaqueness. On the one hand, this implies that it is very difficult to get a more hands-on understanding of the synthesis process[6]. On the other hand, the usage of a set of deep learning algorithms significantly contributes to the solution's excellent data privacy properties. As the original data is only used for training purposes, it is impossible to trace back the original values while looking at the synthesized data. This is very different when compared to traditional data anonymization and encryption techniques, where original information can often be extracted if the potential attacker knows the anonymization procedure or the encryption key. With the synthesized data, this is simply not possible.

Figure 1: Main steps of a synthesis process.

[5] For example, for banks' balance sheet data we impose total assets equal to total liabilities plus equity.

[6] The specific underlying set of algorithms is a proprietary knowledge of the contractor. Even if this information were available to us, it would still be very difficult to gain further insight into the precise inner workings of the model due to the opaqueness of the likely-used methods.



4. Fidelity of the synthesized data

To explore the fidelity of the generated datasets with respect to their original versions, we conduct a series of tests and comparisons. The analysis is two-fold: In the first part, it aims to assess the *statistical properties* of the synthesized data, its distribution, in-sample correlations and general structure when compared to the respective originals. In the second part, we use one of the synthetic datasets in place of the original to judge its usefulness in *modeling applications*.

4.1. Statistical properties of the synthesized data

Dataset 1: Banks' balance sheets

We compare the balance sheets data of banks generated using the synthesis methodology with the real sample. To do so, we try to answer the following questions:

- Are the total aggregates preserved?
- Selecting different banks in different places of the total assets distributions, would they maintain same ratios between selected variables?
- By looking at the univariate distribution and at the bivariate scatter plots, would the synthetic data be similar to the original ones?
- Is the generated database preserving the characteristics and features of the original database?

Results in Table 2 indicate that the new databases, especially the one where rules are not implemented, fail to maintain the total, resulting in minor differences from the original set of data. Despite this limitation, the total sum remains relatively close to that of the original value. Depending on the use case, one could attempt to implement stringent rules during the synthesis process to ensure the preservation of total aggregated thereby enhancing the trustworthiness of the new sample for analytical purposes.

Table 2: Total aggregates for a selection of variables.

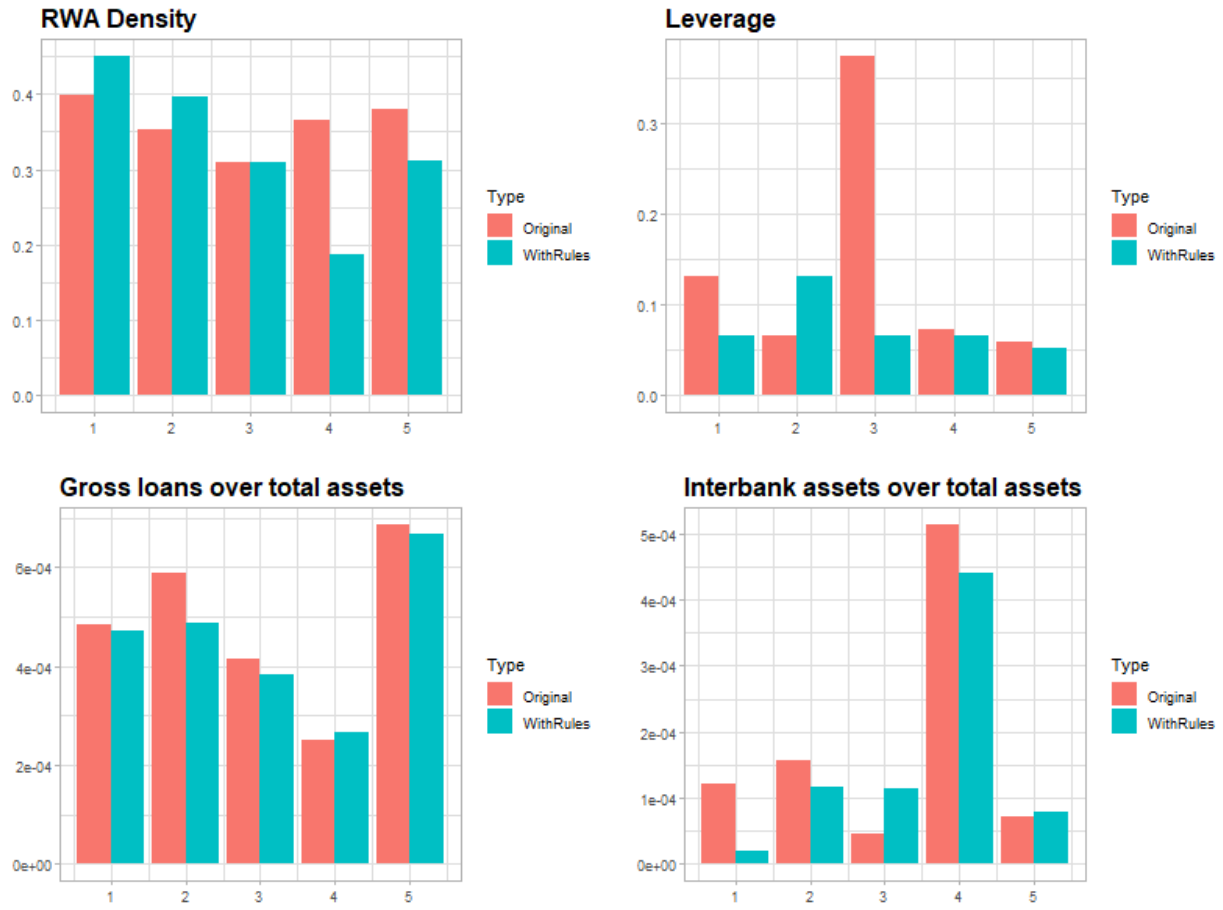
Type	TA	TL	Equity	Tier1	RWA	Gross loans	IB assets
Original	13.13	11.49	1.64	1.97	6.24	0.01	0
With rules	11.92	10.43	1.49	1.61	4.46	0.01	0
No rule	21.23	18.91	2.50	2.64	8.02	0.01	0

We proceed focusing on the synthesized database with enforced algebraic equation rules (to maintain the desired deterministic relationships between the variables) and we compare specific banks within the two datasets. We look at the largest banks in terms of total assets as well as those ranking at the 90th and 75th percentiles, and median bank for each dataset (Table 3). In addition, Figure 2 shows a comparison of some key ratios for the top 5 institutions, specifically: risk weighted assets over total assets (*RWA density),* capital over total assets (*leverage*), gross loans over total assets and, interbank over total assets. While preserving the original scale, results indicate that the synthetized variables differ from the original ones, making it challenging, if not impossible to exactly identify the institutions behind. For example, one can see how the Tier 1 ratio differs for the largest bank, making it more capitalized and less risky (see also the barplots, where a bank among the top 5 shows a very high leverage ratio).

Table 3: Single banks.

	Type	TA	TL	Equity	RWA	Tier1 ratio
Largest	Original	458.89	428.13	30.76	206.27	0.151
	With rules	438.16	380.80	57.36	174.62	0.289
90th	Original	15.83	13.72	2.11	6.95	0.274
	With rules	17.16	15.82	1.33	7.30	0.140
75th	Original	4.00	3.75	0.25	1.37	0.174
	With rules	3.64	3.35	0.29	1.18	0.255
Median	Original	1.44	1.16	0.27	0.85	0.320
	With rules	1.31	1.23	0.08	0.50	0.162

Figure 2: Top 5 banks for the original data set are represented in red, while for the synthetic dataset, where rules are applied, they are highlighted in blue.



Hence, we compare the univariate distribution of equity and the bivariate distribution for assets versus equity, in both the original and synthetic databases. The comparison is made by dividing the sample in three groups: the top 25% of banks (Figure 3), the banks in the second and third quartile (Figure 4),[7] and the banks in the bottom 25% (Figure 5). The distribution of equity is quite similar, but when one look at the bivariate (i.e. scatterplots) there is a clear shift toward the left for some of the largest banks.

[7] We are referring to the banks that fall within the 25% and 75% range in terms of assets.

Figure 3: Top 25% banks for the original data set are represented in red, while for the synthetic dataset, where rules are applied, they are highlighted in blue.

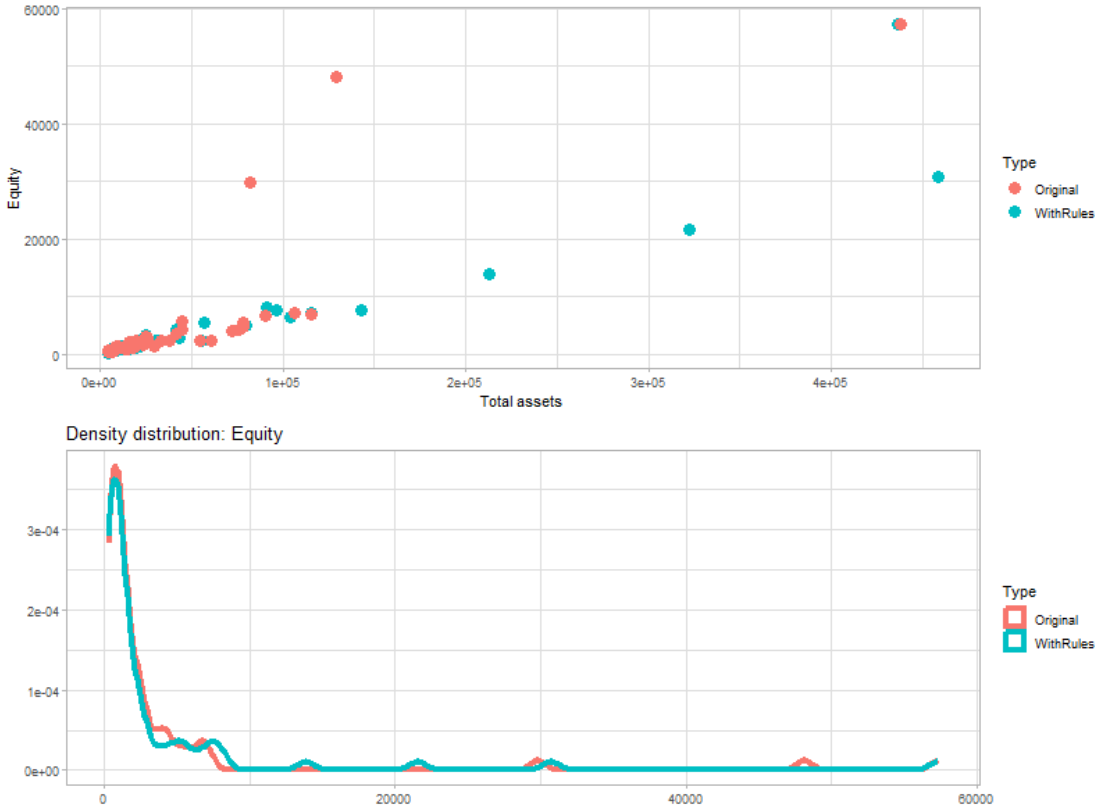


Figure 4: Banks in the inter-quantile range of the original data set are represented in red, while for the synthetic dataset, where rules are applied, those banks are highlighted in blue.


Equity
600
400
200
0
1000
2000
3000
4000
Total assets
Type
Original
WithRules

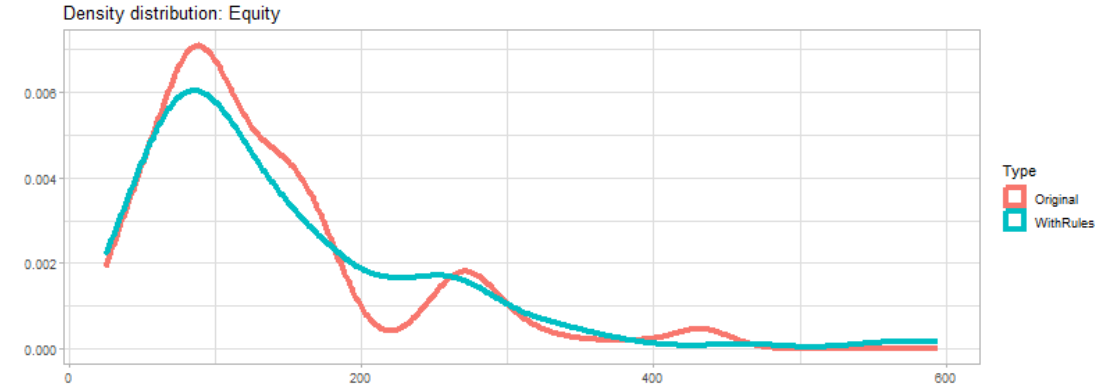
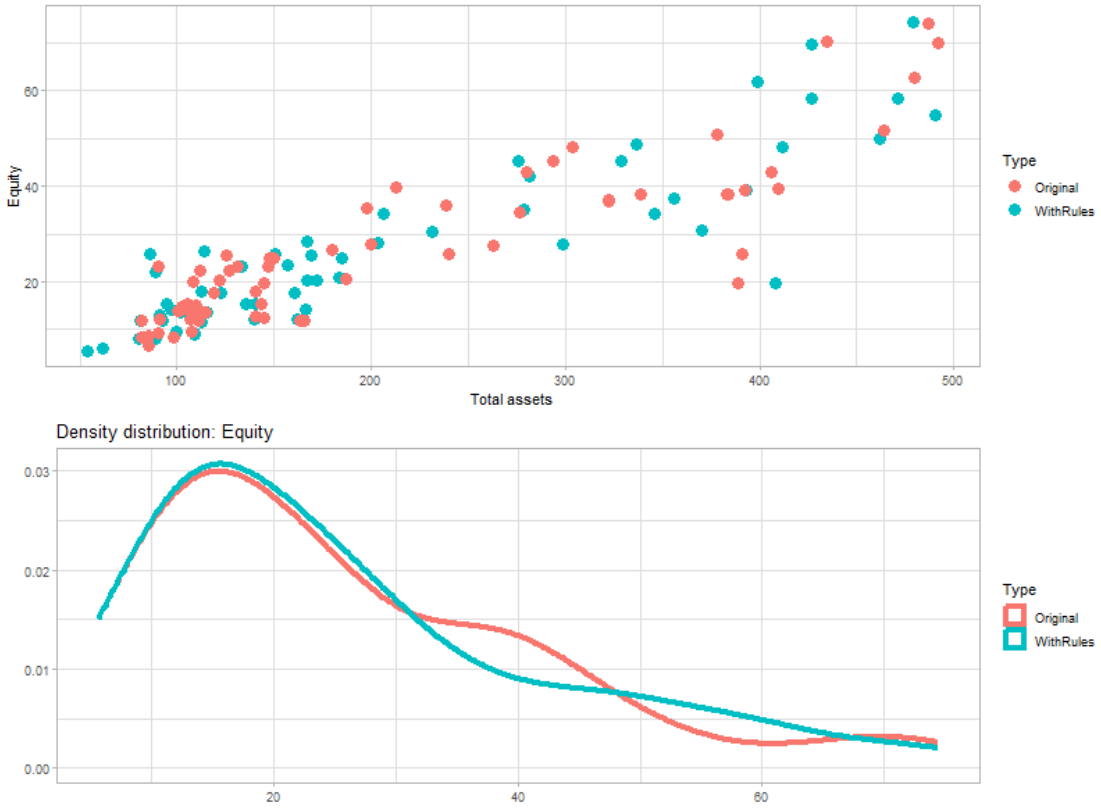

Density distribution: Equity
0.006
0.004
0.002
0.000
0
200
400
600
Type
Original
WithRules

Figure 5: Banks in the bottom 25% inter-quantile range of the original data set are represented in red, while for the synthetic dataset, where rules are applied, those banks are highlighted in blue.



Finally, to detect whether synthetized data contain unintentional unusual numbers or patterns in the main variables of interest, we rely on the Benford's law, a mathematical tool widely used in finance to detect anomalies. This law assumes that in many occurring collections of numbers, the leading significant digit is likely to be small. As bank' balance sheet data is expected to follow this distribution, we statistically measure conformity of the distributions to the expected pattern according to Benford's law. The chi-square ($\chi 2$) test is often used for this purpose, essentially it is a statistical test used to determine whether the distribution of data is significantly different from what is expected. If the calculated $\chi 2$ value exceeds a specific threshold (critical value),[8] it indicates that the actual distribution differs from the theoretical distribution. If the value is less than the threshold, then the distribution follows the expected pattern. We test each variable in the original and synthetized database (see Figure 6). As expected, the ratios do not satisfy the law in any of the dataset. For variables in level, the original dataset always respects the law, whereas the synthetized dataset falls short for net loans, gross loans, risk weighted assets, Tier 1 capital, and Equity. Of course, this raises concerns about the suitability of the synthetized dataset for

[8] The critical value for the chi-square test is determined by the level of significance we want and the degree of freedom. The significance level is the predetermined threshold (usually 0.01) used to determine whether the differences between the two distributions are statistically significant.

conducting bank specific analyses, as these variables appear to deviate from actual balance sheet figures.

Figure 6: Result of accordance to Benford' law.

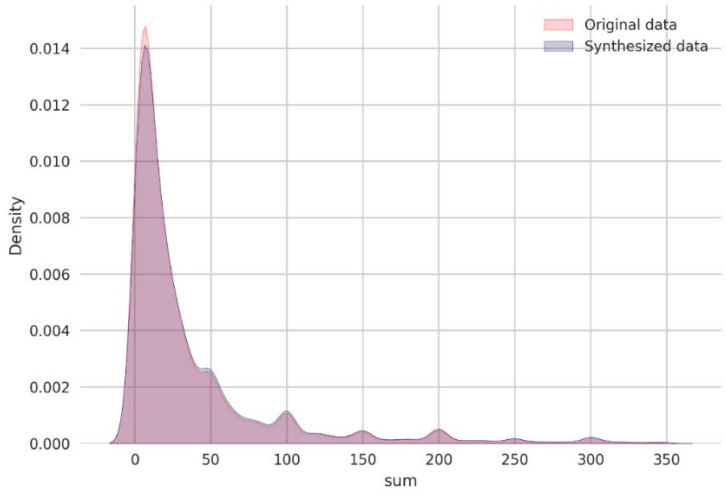
Variable	Original	With rules
Total assets		
Total liability		
Equity		
Tier 1 capital		
Risk weighted assets		
Gross loans		
Net loans		
Loan loss reserves		
Interbank assets		
Derivatives		
Securities and investments		
Other earning assets		
Total earning assets		
Fixed assets		
Total other earning assets		
Equity over total assets		
Risk weighted assets over total assets		
Gross loans over total assets		
Interbank assets over total assets		
Tier 1 ratio		

***Note**: green cells indicate that Benford' law is satisfied; red cells indicate that Benford' law is not satisfied (99% confidence)*

Dataset 2: Retail bank account transactions

We compare the distribution of the retail account transaction amounts between the original and the synthesized dataset. Figure 7 shows the polynomial interpolations used to approximate the density functions for the relevant distributions. The original distribution is skewed, with most of its mass in the area of transactions under 100 EUR – as expected in retail account payments such as groceries. In comparison, there are also concentrations of mass in higher valued transactions which correspond to commonly used manual transfer amounts (such as exactly 100 EUR, exactly 200 EUR, etc.). The Figure shows that the synthesis algorithm is exceptionally good at mimicking this original distribution, including the "regular" transaction amounts.

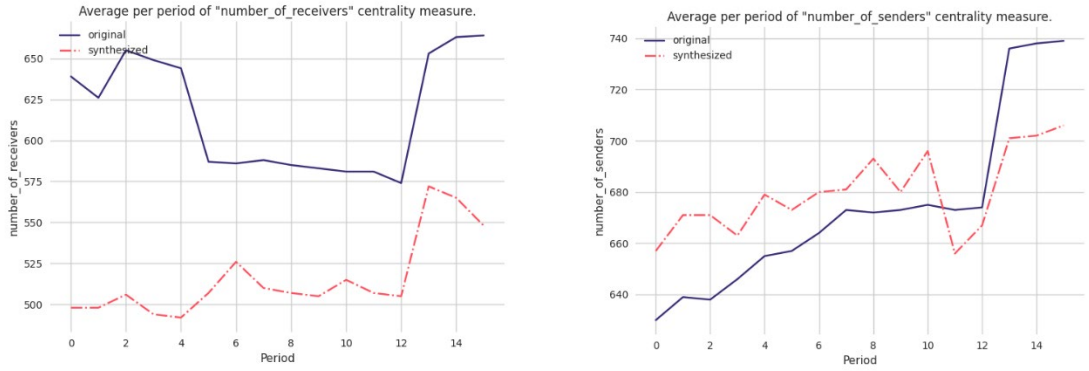
Figure 7: Density functions of the transactions in the original and synthesized database.



Dataset 3: Interbank lending

We use the dataset on Austrian interbank lending to assess whether the synthesis process can replicate network topology reasonably well. Similar datasets are usually used to show the relative importance of the respective banks in the whole system. Figure 8 shows the number of banks that act as lenders and borrowers, respectively, in each period. We find that the synthesized data generally falls within 20% of the original data in both cases. However, there is a consistent overrepresentation (or underrepresentation) of this statistic in the synthesized data.

Figure 8: Density functions of the transactions in the original and synthesized database



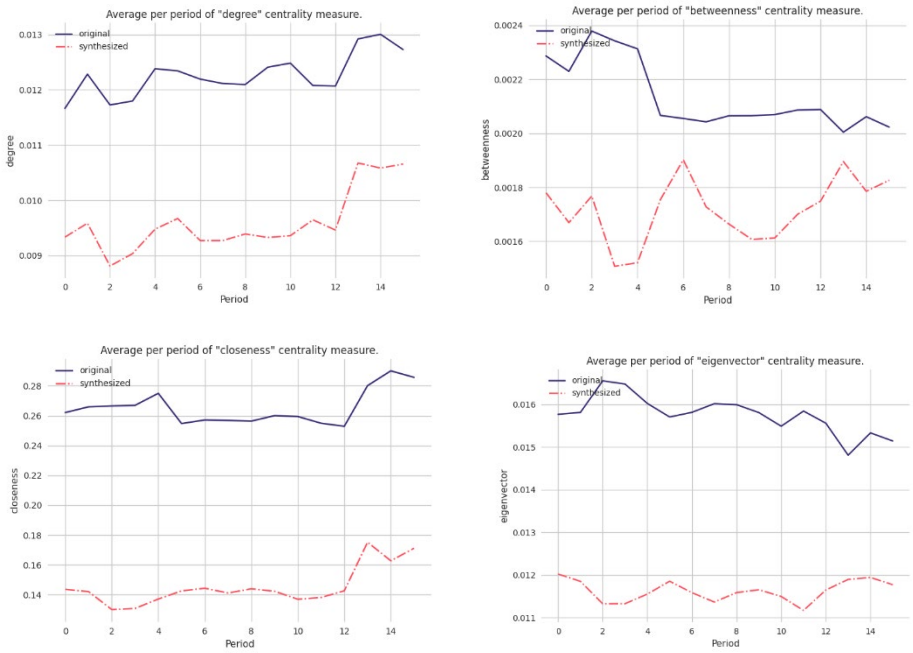
In addition, we explore how the synthesis process preserves various different network centrality measures. For this purpose, we use the following measures:

- *Degree centrality measure* – a degree of a bank is equal to the sum of its distinct borrowers and lenders.

- *Closeness centrality measure* – average shortest distance to all other banks in the network (e.g. if bank A lends directly to bank B, then A's distance to B is equal to 1. If it does not, but instead A lends to C and C lends to B, then A's distance to B is 2 etc.).
- *Betweenness centrality measure* – total number of times that a bank finds itself on a shortest path between all other pairs of banks in the network.
- *Eigenvector centrality measure* – measures the relative "influence" of a bank in the network. If a bank is connected to many banks with a high eigenvector centrality, its own eigenvector centrality is also high and vice versa.

We plot comparisons of these centrality measures in Figure 9. We observe that in all cases, the synthesized data is more uniform / less centralized by a degree of 20% to 50%, depending on the particular centrality measure. In other words, the synthesis process does not effectively replicate the topological network structure, a feature that is crucial when analyzing interbank networks.

Figure 9: Centrality measures for network topology



4.2. Use of synthetic data in a micro-simulation portfolio model

To evaluate the impact of using synthetic data as opposed to real data, we conducted an analysis using the Systemic Model of Banking Originated Losses (SYMBOL).[9] This is a micro simulation portfolio model, based on bank level data, which simulates crisis scenarios where individual banks may default due to their probability of default and level of actual capital. Simulations results are used to approximate the EU loss distribution for the banking sector as well as the size of potential losses affecting the system in case of a systemic crisis. By running the model with both types of datasets, we are able to identify any notable difference and assess the reliability of generated synthetic data.

Table 4 shows that the size of EU losses is approximately 0.3% of total assets, with not major difference when using the synthetic data. Notably, the introduction of specific rules for generating the new dataset appears to align the final results more closely with those obtained using the original dataset. This confirms our previous finding that the implementation of constraints during the data synthesis provides more reliable outcomes that closely mirror the patterns and characteristics observed in the original dataset. Note that despite the fact that synthetized data (with rules) do not satisfy the Benford's law, still they may be useful for a microsimulation model.

Table 4: Bank losses using the original and synthetized database

Type	Losses Over total assets
Original	0.347%
No rule	0.278%
With rules	0.359%

5. Tests performed to check for confidentiality issues

Traditionally, there exist different anonymization techniques which ensure that sensitive information is not released publicly. For instance, these might include removing entire variables (such as names or addresses) or replacing them with hashed values according to a specific algorithm and key. Sometimes, even numeric values are sensitive, which forces the data administrators to either perturb them with random noise or to use a deterministic data transformation which itself acts as a key to the anonymization. In other instances, administrators tend to remove specific observations that correspond to easily identifiable outliers.

[9] De Lisa,R., Zedda, S.,Vallascas, F.,Campolongo, F., andMarchesi,M. (2011). Modelling deposit insurance scheme losses in a Basel 2 framework. Journal of Financial Services Research, 40(3):123–141.

When compared to these individual anonymization techniques, the data synthesis process is fundamentally different. Instead of keeping the original observations and trying to remove the identifiable marks, it "decomposes" the whole dataset into a set of parameters. These parameters can be thought of as basic building blocks which attempt to characterize the original dataset without saying anything about its individual observations. For instance, a simple numerical variable could be characterized by its first two moments – its mean and variance. Afterwards, one could randomly sample from a Gaussian distribution with these exact parameters and create a synthesized version of the original data. Irrespective of the dataset, these two parameters carry only a very limited information about the original data. By looking at them, it is not possible to infer the original individual observations.

Data synthesis decomposes the original data in a similar fashion, but instead of only using two parameters per variable and a very simple probabilistic distribution, the model is much more complex. Admittedly, we have not managed to gain access to the Synthesized's precise algorithms and code, as this is their proprietary knowledge. Nevertheless, we understand from working with the SDK that the process leverages a popular machine learning package Tensorflow, which is mostly used for applications of deep learning. The synthesis process therefore likely runs a machine learning model involving a set of neural networks. In the end, it works exactly as in our simple example with mean and variance, except with a much wider range of parameters that are present in the model.

From a privacy standpoint, this is excellent news. As opposed to traditional anonymization, there is virtually no risk of inferring the precise original data, simply because this data no longer exists in the synthesized version. There is no key, no algorithm, no inverse function one could use to get back the original data. The best one can do is to assume the role of a potential attacker and to see how "close" to the original data one could potentially get. This approach is largely dependent on the precise business case and logic. Nevertheless, we will look into these potential attacks in this section.

5.1. Example of a potential attack by a naive hacker

This session provides insights into the robustness of generated synthesized data against potential attacks that try to figure out confidential information.

For instance, we consider a scenario where a hacker seeks to extract the original value of a specific variable, such as 'gross loans'. We suppose the hacker has access to the synthesized database, knows the value of assets in the real dataset, and is well aware of the high level of correlation between the two variables. By using a simple linear regression model on the synthesized dataset, we establish the relationship between assets and loans. This relationship is then applied to the actual data on assets held by the hacker in order to derive the real value of loans. The analysis is developed using:

a. Full sample

b. Three different subsamples differentiating banks in terms of total assets (banks in the top 25%, banks in the 25%-75% quantiles, banks in the bottom 25%)

Table 5 presents the goodness of fit (R^2) in the real dataset, namely the ability of total assets to explain the size of gross loans. A high value, such as in this case (75% - 96%), indicates a strong correlation between the two variables and thus a better fit. Thus, we apply the estimated regression coefficients to the real value of assets to retrieve the gross loans. Results in Figure 10 show that estimated loans highly differ from the real ones especially when medium and small banks are considered. In Table 6 we report the share of estimated values whose difference with respect to the observed ones fall within specified ranges (1%, 5%, 10% and 20%). For example only 1% of observations falls within 1% relative difference, when the full sample is considered, and only 2% with different regressions according to banks' size.Therefore, it could be inferred that for a hacker, it would be extremely challenging to derive the real bank data solely by relying on the synthetic dataset, even with partial information on the original database.

Table 5: Regression of gross loans with respect to total assets.

	Full sample	*Top 25%*	*25%-75%*	*Bottom 25%*
R^2	*96%*	*96%*	*87%*	*75%*

Figure 10: Percentage difference between the original and estimated dataset, per type of banks. Dotted lines refer to 10% error.

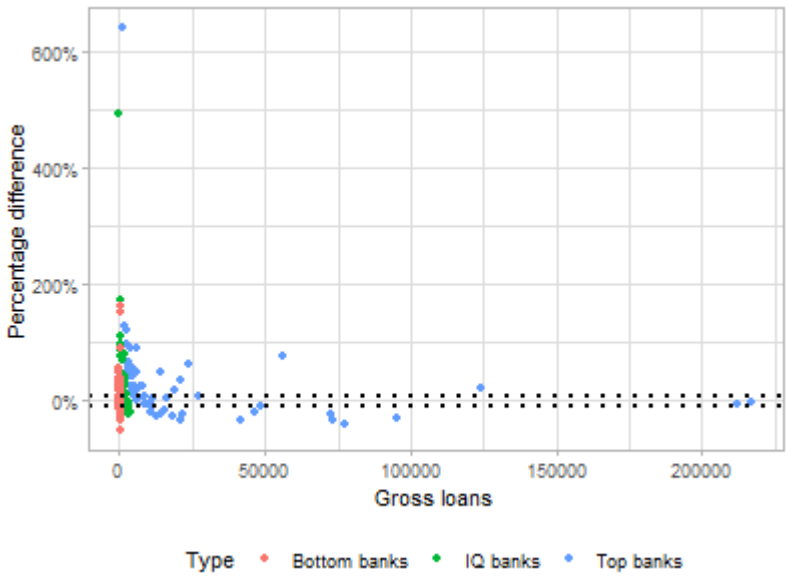


Table 6: Percentages of estimated values.

Percentage of estimated values with different ranges

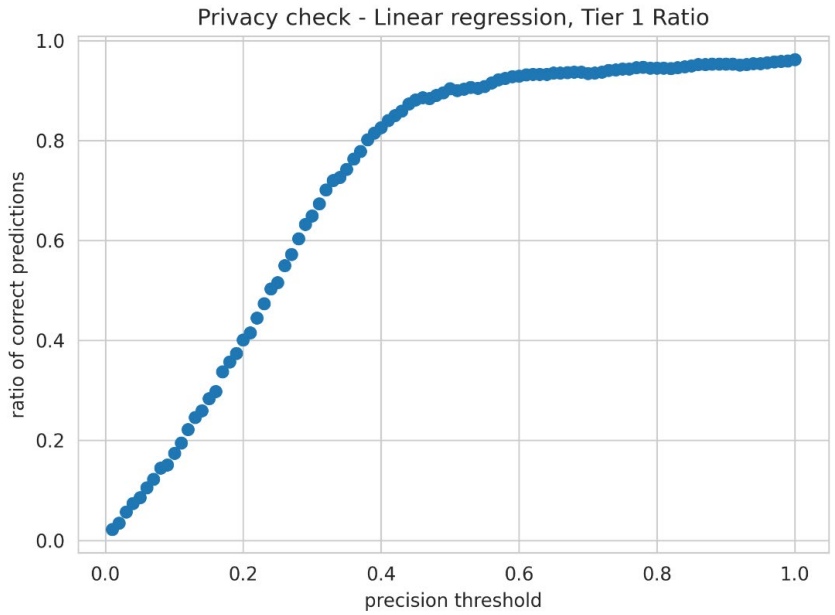
	1%	5%	10%	20%
Full sample	1%	8%	12%	24%
By size	2%	13%	30%	62%

In a second scenario, we assume that the attacker has the full original dataset except Tier 1 capital, risk weighted assets the Tier 1 ratio. Their goal is to estimate the original banks' Tier 1 ratios by using the information from the synthesized data. The procedure is as follows:

1. Develop a model to explain the Tier 1 ratio using all other variables in the dataset and then apply it to the synthesized dataset.
2. Use the estimated coefficients to predict the Tier 1 ratio in the original data.
3. Select an accuracy threshold. If the estimated value is within this percentage distance from the original value, we say that the attacker managed to infer the original value successfully.

In Figure 11, we plot the relationship between the accuracy threshold level and the ratio of observations for which the attacker was successfully able to infer the hidden variable. We assume an unsophisticated attacker and use the simplest possible model for inferring the value – a linear regression. Therefore, these results should be treated as a sort of best case scenario boundary. Nonlinear and more sophisticated models might allow the attacker to correctly identify a higher number of the hidden observations. For this particular dataset and for this particular hidden variable, we can see that for a +/- 5% accuracy, the attacker is able to correctly infer less than 10% of the original data.

Figure 11: Distribution of correct Tier 1 ratio predictions.



6. Conclusion

The goal of this report is to present the tests and checks performed to evaluate the accuracy, anonymization and confidentiality of the synthetic data generated by a software package provided by Synthesized.

There is evidence that newly generated data successfully replicate the main patterns of the original data. While univariate distributions appear to overlap quite well, bivariate distributions reveal some differences. To ensure that these divergences do not hinder the potential use of synthetic data, we use the new dataset in a micro-simulation portfolio model that generates losses in the banking sector under different crisis severities. The results indicate that the aggregated losses do not significantly differ when using the dataset of synthesized banks that impose specific rules as inputs. These finding might suggest that while the divergences allow to preserve effective anonymization, they do not affect the results of a quantitative model using this data as inputs.

Finally, the report tries to assess whether a synthesis algorithm preserves the confidentiality necessary to allow a National Competent Authority to share their confidential information in the Data Hub. As opposed to traditional anonymization techniques, the synthesis process makes any potential attack virtually impossible. This is because instead of keeping an anonymized version of the original data, the synthesis strips it to a limited set of parameters which is then used to build up a completely new dataset with similar statistical properties. This makes any identification of individual original observations impossible, because they are simply no longer there. In the best case, one can try to estimate how “close” a potential attacker might get to the original information, provided we are dealing with a cardinal variable. For a reasonable precision level, we show that the attacker is able to correctly attribute only a very limited portion of the original data.

Getting in touch with the EU

In person

All over the European Union there are hundreds of Europe Direct centres. You can find the address of the centre nearest you online (european-union.europa.eu/contact-eu/meet-us_en).

On the phone or in writing

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

— by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),

— at the following standard number: +32 22999696,

— via the following form: european-union.europa.eu/contact-eu/write-us_en.

Finding information about the EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website (european-union.europa.eu).

EU publications

You can view or order EU publications at op.europa.eu/en/publications. Multiple copies of free publications can be obtained by contacting Europe Direct or your local documentation centre (european-union.europa.eu/contact-eu/meet-us_en).

EU law and related documents

Science for policy

The Joint Research Centre (JRC) provides independent, evidence-based knowledge and science, supporting EU policies to positively impact society

EU Science Hub
joint-research-centre.ec.europa.eu

Publications Office
of the European Union