ECONOMETRICS
EKONOMETRIA
Advances in Applied Data Analysis

Econometrics. Ekonometria.
Advances in Applied Data Analysis
Year 2024, Vol. 28, No. 2
ISSN 2449-9994
journals.ue.wroc.pl/eada

Synthetic Financial Data: A Case Study Regarding Polish Limited Liability Companies Data

Aleksandra Szymura

Wroclaw University of Economics and Business, Poland

e-mail: aleksandra.szymura@ue.wroc.pl

ORCID: 0000-0002-9009-3655

Quote as: Szymura, A. (2024). Synthetic Financial Data: A Case Study Regarding Polish Limited Liability Companies Data. *Econometrics. Ekonometria. Advances in Applied Data Analysis*, *28*(2), 1-17.

DOI: 10.15611/eada.2024.2.01

JEL: C45, C80, G30

Abstract

Aim: The aim of this article was to present and evaluate the concept of synthetic data. They are completely new, artificially generated data, but keep the statistical properties of real data. Due to the statistical similarity with real data, they can be used instead of them. This action allows data to be shared externally while guaranteeing their privacy.

Methodology: New datasets were generated based on financial information about Polish limited liability companies, which come from the Orbis database and refer to 2020. To create synthetic data, it was decided to use generative models: CTGAN (based on GAN architecture) and TVAE (based on autoencoders). Lastly, the synthetic data were compared with the real ones in terms of statistical properties (e.g. shape of distributions, correlations etc.) and their applicability in data analysis (the PCA method).

Results: The Overall Quality Score was higher for the data generated by TVAE, but after examining the results in more detail, it was seen that the data generated by CTGAN had a better quality in terms of keeping the statistical properties of the real data. Comparing the results of the PCA method, TVAE was better than CTGAN. In addition, the TVAE method was less time-consuming than CTGAN.

Implications and recommendations: Before publishing the synthetic data externally, it is recommended that the data are generated using several algorithms, evaluating their final results and finally selecting the best option. This action enables the resulting dataset to be of the highest quality. In further research, it is proposed that other algorithms are tested (e.g. CopulaGAN or TableGAN), in an attempt to deal with some of the realistic data problems that were missed in this analysis, such as missing values (the work was carried out with a complete dataset). Data generated in this study may

be used to build financial indicators, which in turn could be used to construct company assessment models.

Originality/value: Synthetic data help to deal with some of the data limitations, such as data privacy or scarcity. Due to their statistical similarity with real data, it is possible to use them in advanced machine learning methods instead of real datasets. Analysis on high quality synthetic data allows conclusions similar to analysis on real data to be achieved, while retaining privacy and without publishing sensitive data to third parties.

Keywords: synthetic data, generative models, financial data, CTGAN, TVAE

1. Introduction

Nowadays, given the recent growth of technology, data are considered as the most valuable resource, and assessed as having the same value as gold (Pathare et al., 2023). Many advanced systems or models rely on data (Lu et al., 2023). Every day people leave a digital footprint by sending messages, generating content on social media, logging on to many platforms or doing online shopping. All of this information is collected in dedicated databases by the private sector and/or government. According to the Statista report, the amount of data created will exceed the level of 180 zettabytes by 2025 (Statista, 2023). Thanks to the information gained, it is possible to make more accurate and effective decisions that help people in their daily lives or aid companies in their activities. The decision-making process becomes faster. Unfortunately, it also has another side. It raises some limitations, not only on collecting data and processing methods, but also on data privacy. The last aspect deserves special attention. Apart from open data, much of the information is sensitive, such as personal, financial or medical details which require greater protection. Their processing is extremely limited, very often regulated by law, e.g. General Data Protection Regulation (The European Union, 2016) and the latest Artificial Intelligence Act (AI Act), which is in the process of being drafted by the European Commission (European Parliament, 2023). Each data leakage can cause negative consequences. Many companies and institutions do not have enough hardware resources to train advanced machine learning models, so they need to use cloud services. Unfortunately, data privacy restrictions very often do not allow data to be shared externally (Efimov et al., 2020). In addition, sharing datasets enables many researchers to work on a particular problem at the same time. Competition is growing, so the new solutions are better, but they can also be shared simply as examples of data for educational purposes (Rajotte et al., 2022). With all the restrictions, this action has become impossible.

The objective of this article was to present a solution that can help to deal with the above mentioned limitations with data. Here the concept of synthetic data, completely new data that are imitations of real ones, is introduced, which are generated on the basis of already existing real data, keeping their statistical characteristics. In short: brand new, but similar to the real data. It is also possible to generate synthetic data without real data using existing models or expert knowledge (Centrum Nowych Technologii dla Polityk Publicznych NASK-PIB, 2022). Synthetic data generation is better than classic anonymisation methods such as data masking or encryption, because new records, similar to the real ones, are generated. Anonymisation is the removal or replacement of sensitive information or data that can be identified, and for this reason it may not guarantee full privacy.

The synthetic data concept is also strongly linked with the topic of data imputation. It is not only possible to generate completely new whole records, but also to generate single values to impute missing values (Wang et al., 2021). Imputation methods can be classified as discriminative or generative (Neves et al., 2022). Advanced methods, based on deep learning, as mentioned in this article, namely autoencoders or generative adversarial networks (GANs), belong to the generative category (Yoon et al., 2018). In this work, the problem of generating the whole datasets was addressed.

The aspect of the statistical similarity between synthetic and real data is discussed in this paper. Some evaluation metrics were calculated and their usefulness in data analysis was tested. The paper is structured as follows: in the second section the result of the literature review on synthetic data and generative models is presented. In the third section, the methodology is shown. The fourth section is a presentation of the dataset and all the results obtained. The final section contains conclusions and proposals for further research directions.

2. Literature Review

The concept of synthetic data has been present in the literature for years, but has become more popular recently due to the development of generative artificial intelligence. Synthetic data are the response to several limitations with data, allowing data scarcity to be managed. In fact, many real-world datasets are affected by this problem, and many of them are heavily imbalanced. Szymura (Szymura, 2022) used synthetic data generated by SMOTE (Synthetic Minority Over-sampling Technique) to increase the minority class. In the article (Khaemba et al., 2023), Khaemba and other researchers tackled the topic of data generation in response to the lack of datasets for Agetech. Hameed and Alamgir (Hameed & Alamgir, 2022) used also SMOTE and generative models (generative adversarial networks and variational autoencoders) to cope with the problem of small and imbalanced datasets. The problem of data scarcity is also linked to that of missing values in datasets. Imputing data using new, synthetic data avoids removing incomplete records (Neves et al., 2022). In their paper (Wang et al., 2021), the authors used pseudo-label conditional generative adversarial imputation networks (PC-GAIN) to deal with incomplete data problems. Shahbazian and Trubitsyna in their study (Shahbazian & Trubitsyna, 2022) proposed a novel method for handling missing data – DEGAIN – which is an improved version of generative adversarial imputation networks (GAIN) proposed by Yoon, Jordon and Van Der Schaar (Yoon et al., 2018). Compared to GAIN, the deconvolution concept was added to DEGAIN. In this study, DEGAIN performed better than GAIN.

Synthetic data can be used instead of real data to training models. Muñoz-Cancino and others (Muñoz-Cancino et al., 2022) made an attempt to evaluate creditworthiness assessment models based on real-world data and synthetic data generated by CTGAN and TVAE. The authors concluded that, despite results that were not fully satisfactory, by protecting the data privacy, such an action could boost cooperation between financial institutions and academia. Efimov and others (Efimov et al., 2020) used GAN models to generate financial data and checked the usefulness of this data in machine learning models, and concluded that synthetic data were really close to the real data. Evaluation metrics for supervised models based on real data were slightly higher than for synthetic data. In addition, they used an unsupervised learning algorithm, t-SNE. The synthetic data reproduced the clusters based on the real data with good accuracy in terms of shape and density. Yilmaz and Korn (Yilmaz & Korn, 2022) used generative adversarial networks (RCGAN, TimeGAN, CWGAN and RCWGAN) to generate individual electricity consumption data. They concluded that models based on GAN architecture are able to produce realistic, good quality synthetic data. Sivakumar et al. proposed in their paper (Sivakumar et al., 2023) GenerativeMTD which uses VAE-GAN-like architecture in order to generate small datasets. The authors also presented the performance of classification and regression methods on synthetic data generated by other methods such as VEEGAN, TableGAN, TVAE and CTGAN, concluding that GenerativeMTD outperforms all these methods in terms of data quality and privacy. Models based on GAN architecture are also used to generate stock market data. In their paper (Li et al., 2020) the authors used Stock-GAN, among others, and found that the synthetic data were similar to the actual market data. In another article (Carvajal-Patiño & Ramos-Pollán, 2022) generative models were used to generate new datasets to help build trading strategies.

Synthetic data is a very popular concept in the medicine and biometrics domain, where many data are sensitive. Murtaza and other researchers (Murtaza et al., 2023) presented a review of the usage of synthetic data in the medical domain preserving data privacy. Another study (Choi et al., 2017)

proposed a new approach to creating new patient records, which was a method based on generative adversarial network – medGAN. Beyond its usefulness in solving analytical problems, they also concluded that, using medGAN, the risk of attributed values disclosure was reduced. Bamoriya et al., used DSB-GAN to generate biometric data (Bamoriya et al., 2022). This method is based on convolutional autoencoders (CAE) and deep convolutional generative adversarial networks (DCGAN). The authors found that DSB-GAN performed very well in image generation. The images generated by DSB-GAN were clear and complete compared to those generated by other methods. Other authors (Karbhari et al., 2021), to deal with data scarcity, tried to generate chest X-rays using an auxiliary classifier generative adversarial network (ACGAN). Then, they used convolutional neural networks (CNN) to detect Covid-19 using two datasets. The first contained the original data combined with synthetic data, and the second only the original data; all the models achieved an AUC value greater than 0.98.

3. Methodology

Based on the results of the literature review, the author decided to use two methods to generate synthetic data: the conditional tabular generative adversarial network (CTGAN) and variational autoencoders adapted to tabular data (TVAE). Both are briefly described in this section. In addition, the methods for evaluating synthetic data are also outlined.

3.1. CTGAN

Generative adversarial networks are very often applied to produce new data. One of the models based on GAN architecture is the conditional tabular generative adversarial network (CTGAN), first proposed in (Xu et al., 2019). Thanks to *mode-specific normalisation*, it can cope with non-Gaussian and multimodal distribution (Xu et al., 2019). In previous models, such as TableGAN, min-max normalisation was used to normalise continuous variables to [–1,1] values (Bourou et al., 2021). CTGAN can also overcome the problem of different data types and the imbalance of categorical variables (Centrum Nowych Technologii dla Polityk Publicznych NASK-PIB, 2022; Inan et al., 2023).

3.2. TVAE

Autoencoders are unsupervised learning methods that are used in particular to deal with two analytical problems: dimensionality reduction and synthetic data (Muñoz-Cancino et al., 2022). Autoencoders are composed of two parts, an encoder and a decoder. The first part, the encoder, transforms input data into a latent space, whilst the second part, the decoder, transforms data from a latent space into output data. Variational autoencoders are based on and are an extension of autoencoders. They can help overcome the limitations of autoencoders (Figueira & Vaz, 2022). This method was first proposed in (Kingma & Welling, 2013). Compared to classical autoencoders, variational autoencoders attempt to map input data into a multivariate Gaussian distribution in the latent space, and not to a vector, using an encoder (Figueira & Vaz, 2022; Podolszańska, 2021). The Tabular Variational Autoencoder (TVAE) is a type of variational autoencoders, adapted to generate tabular data (Xu et al., 2019), by using the evidence lower bound (ELBO) loss (Muñoz-Cancino et al., 2022).

3.3. Evaluation Metrics

Before publishing new synthetic data outside, it is necessary to evaluate them – comparing to the real data. Many of the metrics mentioned in this paper come from the Python package – *SDMetrics* (*Synthetic Data Metrics*, 2023), thus their descriptions below are based on the documentation of this package. The following metrics were used to evaluate new datasets:

1. KSComplement: used to compare the shape of numerical, continuous columns. It is based on Kolmogorov-Smirnov statistics and the empirical Cumulative Distribution Function (CDF), and takes the values between 0 and 1. The higher the value, the more similar the marginal distributions of the real and new data are. This metric is computed as 1 minus KS Statistic.
2. BoundaryAdherence: used to evaluate whether the values of the variables from the synthetic data respect the minimum and maximum boundaries of the same variables from the real dataset. The higher the value for a variable, the more values from the new data there are between the minimum and the maximum values from the real data.
3. The Spearman rank correlation coefficient: used to explore the relation between variables in each datasets. Compared to the Pearson coefficient, it is less sensitive to the occurrence of outliers, and takes the values between –1 (strong negative relation) and 1 (strong positive relation). Its absolute value allows the strength of the relations to be determined.
4. NewRowSynthesis: used to assess the novelty of records from synthetic data. The values are from 0 to 1; the higher the value, the more rows in the synthetic data do not match the real data.
5. Descriptive statistics: the mean, minimum value, quartile I, median, quartile III, maximum value and coefficient of variation were used to evaluate the statistical similarity between each variable from the synthetic and real data.

3.4. Principal Component Analysis

Principal component analysis (PCA) is one of the unsupervised learning techniques for dimensionality reduction. This method allows a large set of variables to be reduced to a smaller one, containing only representative variables (Talabis et al., 2015). Correlated variables are replaced by a few unrelated variables, called principal components (Liu, 2022), which are linear combinations of the original variables (Awad & Khanna, 2015). The PCA method enables feature extraction without a significant loss of information.

4. Results

This section presents the results of the analysis. First, the dataset used is presented, followed by the values of all calculated evaluation metrics. Lastly, the usefulness of using synthetic data is tested in machine learning methods using principal component analysis (PCA).

4.1. Dataset

The first step of the analysis was to prepare the dataset. It was decided to generate data based on financial information about Polish limited liability companies. All the required data came from the Orbis database and referred to 2020. The only condition was the collection of a complete dataset, without any missing values. The author is aware of the fact that, in reality, the datasets were not complete, but for the purposes of this work it was decided not to tackle this pre-processing issue.

Firstly, 44 variables for all Polish limited liability companies were downloaded. Only numeric, continuous variables were selected. All the records that contained only *n.a.* (*not available*) or *n.s.* (*no significance*) values were removed. There were 275,407 companies for which at least one single piece of financial information was available. The variables with a high level of completeness of data (> 85%) were selected for the next stage. All the records with one or more missing values were removed. Finally, a sample of 124,284 records and 13 variables was obtained. Table 1 shows the descriptive statistics of the selected variables.

Table 1. Description of variables

Variable	Orbis Database Description	Mean [th USD]	Minimum Value [th USD]	Maximum Value [th USD]	Median [th USD]	Coefficient of Variation [%]
pl_bef_tax	P/L before tax	204.89	-347,517.29	272,680.92	11.71	1391.75
net_income	P/L for period [=Net income]	154.48	-347,447.84	220,167.09	10.11	1643.22
add_val	Added value	842.76	-184,596.10	547,567.31	79.29	679.29
cur_as	Current assets	2119.29	-2885.80	2,401,782.90	224.56	751.81
oth_cur_as	Other current assets	688.27	-2964.56	716,969.71	58.80	854.65
ncur_as	Non-current assets	2514.05	-622.34	3,543,097.80	30.60	879.50
tot_as	Total assets	4633.40	-55.34	5,944,880.70	379.15	728.97
capital	Capital	711.62	-297,119.78	1,235,199.00	13.30	1317.52
sh_funds	Shareholders funds	1917.58	-680,837.85	1,814,005.85	98.71	839.47
ot_sh_funds	Other shareholders funds	1205.97	-924,335.14	1,799,374.78	55.61	1094.36
tot_sh_funds	Total shareholders' funds and liabilities	4634.08	-350.68	5,944,880.70	378.88	728.90
stock	Stock	590.58	-716.53	1,182,202.25	1.06	1106.75
debtors	Debtors	840.52	-295.07	1,182,188.15	66.78	961.56

Source: own elaboration.

4.2. Sample Size vs Overall Quality Score

All the calculations were made using Python packages: *SDV* (Patki et al., 2016) and *SDMetrics* (*Synthetic Data Metrics*, 2023). In the first step it was decided to compare the overall quality of synthetic data made by the two mentioned methods depending on the training sample size – Table 2 shows the results of this stage. The comparison was performed for sample sizes of 10 000, 30 000, 50 000, as well as for the whole dataset. The new datasets had the same size as the input training sample. Given the author's knowledge and experience, and hardware capabilities, the models were trained using modifications of the following parameters: *epochs* and *batch_size*. The value of *epochs* was set as 75 and *batch_size* as 500. The first parameter determines the number of times to train the model, and the second, *batch_size*, specifies the number of samples that will be used to train the model in one epoch. The larger the value of the *batch_size* parameter, the more memory space is required to train neural networks. All the other parameters were set to default values, which is explained in more detail in the *SDV* package documentation[1].

Table 2. The value of Overall Quality Score [%] depending on sample size

Sample size	CTGAN	TVAE
10 000	73.76	85.44
30 000	79.12	88.95
50 000	86.86	89.45
All records	87.90	89.65

Source: own elaboration.

In this example, the larger the sample size, the higher the overall quality of the synthetic data. For the CTGAN method this relation was clearly visible. The values of quality score were between 73.76% and 87.90%. For TVAE, the values of quality score were very close to and higher (between 85.44% and 89.65%), so probably with the next training session (with other parameter values) and generating new

[1] Official website of SDV package: https://docs.sdv.dev/sdv (accessed on 28 October 2023).

data, this relation might not occur. All the evaluating comparisons presented in the next section (4.3) were made on a sample based on all records.

Table 3. Time of training models and generating new data depending on sample size

Sample size	CTGAN	TVAE
10 000	1 min 19 sec	0 min 47 sec
30 000	4 min 38 sec	2 min 6 sec
50 000	8 min 39 sec	4 min 20 sec
All records	22 min 32 sec	8 min 40 sec

Source: own elaboration.

Additionally, the time of training models and generating new data in terms of the sample size were checked – see Table 3 for the results. For each sample size, CTGAN was definitely more time-consuming than TVAE.

4.3. Evaluation of Synthetic Data

Firstly, data quality was checked by comparison of column shapes by comparing marginal distributions. For data generated by TVAE, the average value of the KSComplement for all the variables was higher (0.88) than for those generated by CTGAN (0.84). Figures 1 and 2 show KSComplement metric values by variables.

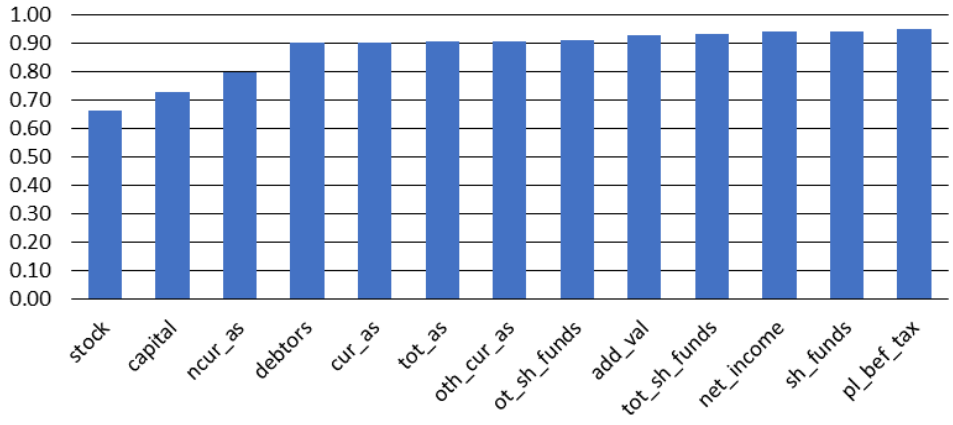

Fig. 1. The value of the KSComplement metric for data generated by TVAE

Source: own elaboration.

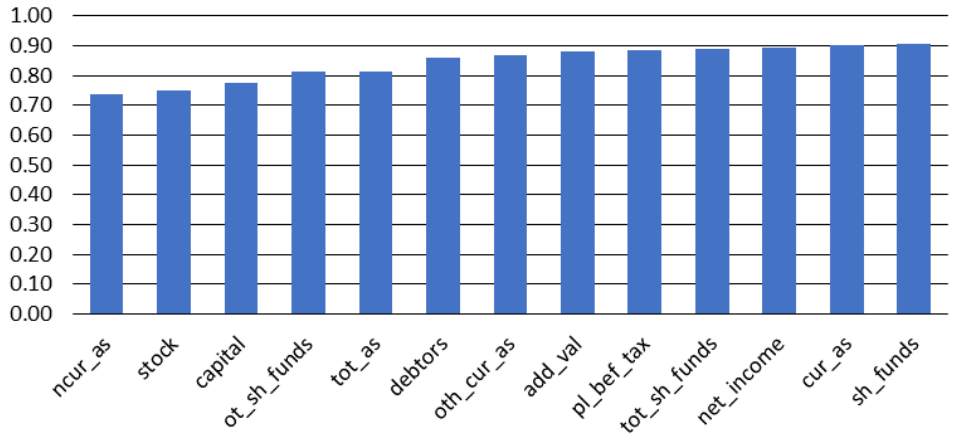

Fig. 2. The value of the KSComplement metric for data generated by CTGAN

Source: own elaboration.

For synthetic data generated by TVAE, the variable *pl_bef_tax* had the most similar shape with real data, for CTGAN – *sh_funds*, however the most divergent shape with real data appeared in variable *stock* for the TVAE method and *ncur_as* for CTGAN. Figures 3-6 show the distribution of real and synthetic data for the mentioned variables. For a better visualisation, the modified logarithmic transformation was used for these variables according to the formula:

$$f(x) = \begin{cases} sgn(x)\, lg|x|, & x <> 0 \\ 0, & x = 0 \end{cases}$$

In Figures 3 and 5 it can be seen that both methods dealt well with mapping the bimodal distribution. In contrast, Figures 4 and 6 show that both methods failed to deal with variables that have a high concentration around a single value, in this example – zero. For the variable *stock* there were approximately 46.14% zero values, and for the variable *ncur_as* – 23.43%. Considering the statistical parameters, these variables were well mapped, but looking at the shape of the distribution, the result was not satisfactory. In some of the mentioned examples multimodal distribution was finally received.

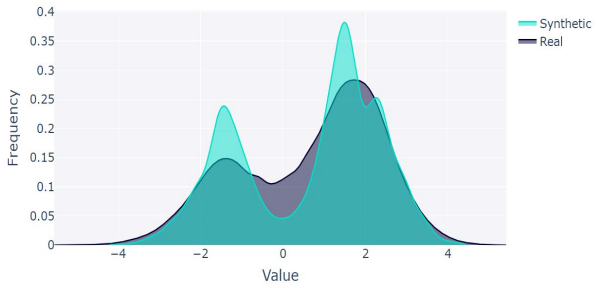

Fig. 3. Comparison of distribution of column *pl_bef_tax* between real and synthetic data generated by TVAE

Source: own elaboration using Python.

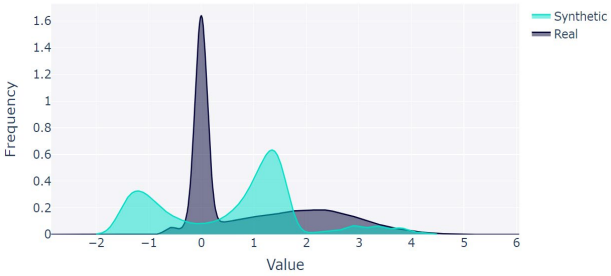

Fig. 4. Comparison of distribution of column *stock* between real and synthetic data generated by TVAE

Source: own elaboration using Python.

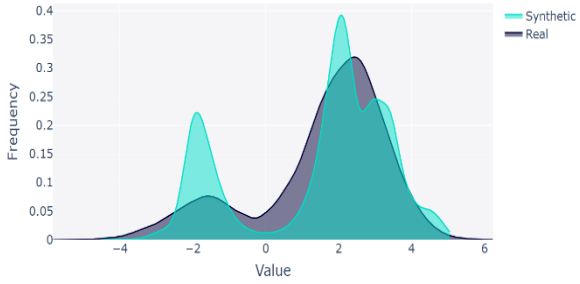

Fig. 5. Comparison of distribution of column *sh_funds* between real and synthetic data generated by CTGAN

Source: own elaboration using Python.

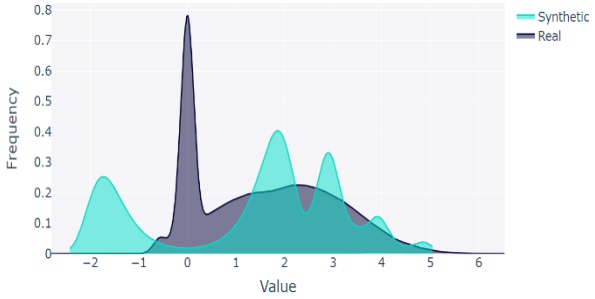

Fig. 6. Comparison of distribution of column *ncur_as* between real and synthetic data generated by CTGAN

Source: own elaboration using Python.

Moreover, the boundary adherence was checked. For all the variables, for both synthetic datasets, the value of these metrics was 1, so all the values respected the minimum and maximum boundaries of the real data, which is also visible in Table 4. All the minimum values for each column in the real data were lower or equal to the minimum values of the corresponding variables in the synthetic datasets. All the maximum values for each column in the real financial data were greater or equal to the corresponding variables in the synthetic datasets.

The next stage of evaluating datasets was the analysis of the correlation between variables in each dataset. It is really important to keep the relations between the variables. To check it, the Spearman rank correlation coefficient was used. Figures 7 to 9 show the correlation matrix. It is worth pointing out that the strength of the relation between the variables was not perfectly reproduced by both methods, however, the direction was maintained. Looking at the *capital* and *stock* variables, it seems that CTGAN dealt with mapping the relations better than TVAE. Unfortunately, both methods do not take into account the adopted financial principles, such as the balance sheet rule, in which the total sum of assets (*tot_as*) must be equal to the sum of total shareholders' funds and liabilities (*tot_sh_funds*). The value of the correlation coefficient between these variables in the synthetic data must be 1, such as in the real dataset.

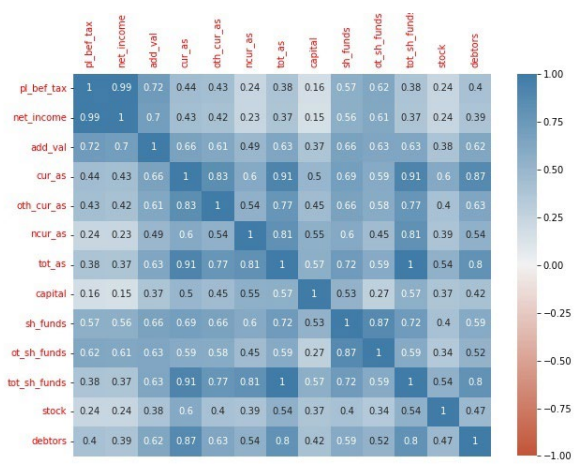

Fig. 7. A plot of the Spearman rank correlation coefficient values between variables from real dataset

Source: own elaboration using Python.

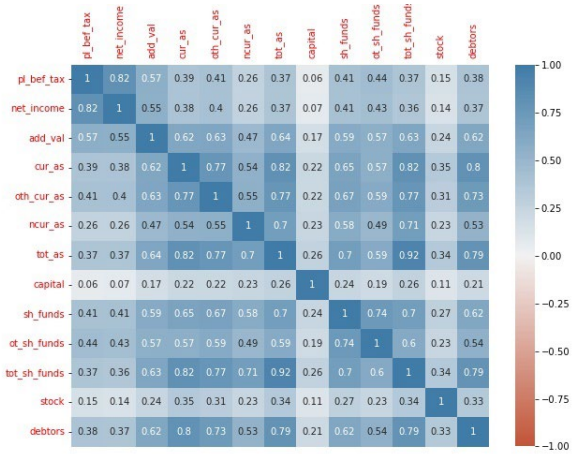

Fig. 8. A plot of the Spearman rank correlation coefficient values between variables from new dataset generated by TVAE

Source: own elaboration using Python.

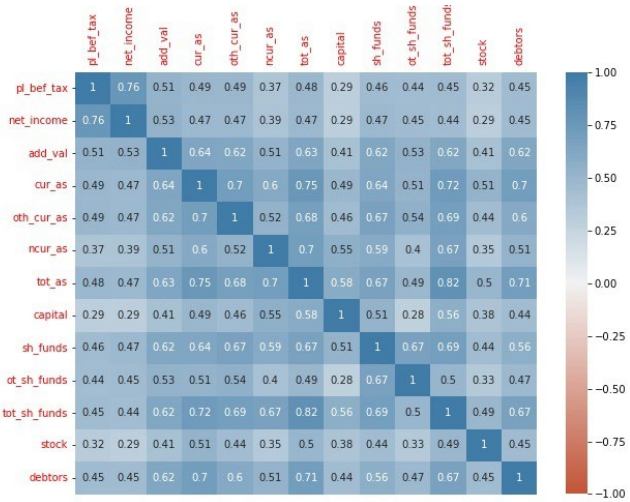

Fig. 9. A plot of the Spearman rank correlation coefficient values between variables from new dataset generated by CTGAN

Source: own elaboration using Python.

The value of the NewRowSynthesis metric was 1 in both cases, which means that all the rows in the new datasets were completely new; none of these matched with the real data. It is a very important aspect to maintain data privacy.

Statistical similarity was checked by comparing the values of some descriptive statistics: mean, minimum value, quartile I, median, quartile III, maximum value and coefficient of variation between real and synthetic data. The results are presented in Table 4. This similarity was computed for all the variables and statistics and the final conclusion is presented in the last column "Greater similarity". Firstly, the values y_{TVAEij} and $y_{CTGANij}$ were calculated according to the formulas:

$$y_{TVAEij} = \left|x_{REALij} - x_{TVAEij}\right|$$

and

$$y_{CTGANij} = \left|x_{REALij} - x_{CTGANij}\right|,$$

where

y_{TVAEij} – the absolute value of the difference between the value of j-statistic for i-variable of real data and the value of j-statistic for i-variable of synthetic data generated by the TVAE method,

$y_{CTGANij}$ – the absolute value of the difference between the value of j-statistic for i-variable between real data and the value of j-statistic for i-variable of synthetic data generated by the CTGAN method.

If the y_{TVAEij} value is lower than the $y_{CTGANij}$ value, then the TVAE method gave a better result than CTGAN, otherwise CTGAN gave a better result. If the y_{TVAEij} value is equal to the $y_{CTGANij}$ value, then both methods gave the same results (marked as "BOTH" in the column "Greater similarity").

Table 4 shows that both methods did not deal with extreme values mapping. The maximum values for all the variables from the real data were significantly higher than for the variables from the synthetic data generated by CTGAN and TVAE. This phenomenon was not so visible in the case of the minimum values, where for some variables from the original dataset they covered the minimum values for the synthetic data. In the real dataset, the minimum values of the variable considered were not greatly different from the other values from that variable compared to the maximum values. The coefficient of variation values (Table 4) were much lower for the datasets generated by both methods than for the real data. It is desirable for the synthetic data to be similar to the real data, also in terms of the occurrence of outliers and extreme values. This was also compared by boxplots presented in Figures 10 to 12. For a better presentation of the data, a modified logarithmic transformation was again applied according to the formula described above in this section. It can be observed that the real values of the variables were significantly higher than for their generated counterparts.

Table 4. Statistical similarity

Variable	Metrics	Real data	Synthetic data – CTGAN	Synthetic data – TVAE	Greater similarity
pl_bef_tax	Mean [th USD]	204.89	275.77	156.72	TVAE
	Minimum Value [th USD]	-347,517.29	-8793.74	-16,042.68	TVAE
	Quartile I [th USD]	-5.85	-17.88	-12.22	TVAE
	Median [th USD]	11.71	8.63	16.80	CTGAN
	Quartile III [th USD]	97.91	112.20	92.43	TVAE
	Maximum Value [th USD]	272,680.92	31,115.52	27,912.86	CTGAN
	Coefficient of Variation [%]	1391.75	554.14	589.34	TVAE

net_income	Mean [th USD]	154.48	158.24	162.18	CTGAN
	Minimum Value [th USD]	-347,447.84	-9813.62	-12,429.48	TVAE
	Quartile I [th USD]	-6.12	-16.70	-12.34	TVAE
	Median [th USD]	10.11	8.99	15.96	CTGAN
	Quartile III [th USD]	84.08	111.90	106.98	TVAE
	Maximum Value [th USD]	220,167.09	14,810.65	15,248.71	TVAE
	Coefficient of Variation [%]	1643.22	502.37	526.45	TVAE
add_val	Mean [th USD]	842.76	795.03	515.19	CTGAN
	Minimum Value [th USD]	-184,596.10	-5725.11	-7709.38	TVAE
	Quartile I [th USD]	8.78	2.58	19.88	CTGAN
	Median [th USD]	79.29	76.37	75.11	CTGAN
	Quartile III [th USD]	383.14	405.38	334.90	CTGAN
	Maximum Value [th USD]	547,567.31	49,805.40	46,547.51	CTGAN
	Coefficient of Variation [%]	679.29	364.00	359.07	CTGAN
cur_as	Mean [th USD]	2119.29	2110.19	1345.38	CTGAN
	Minimum Value [th USD]	-2885.80	-528.08	-1411.90	TVAE
	Quartile I [th USD]	48.69	77.17	75.70	TVAE
	Median [th USD]	224.56	195.28	358.23	CTGAN
	Quartile III [th USD]	916.42	1051.95	1000.56	TVAE
	Maximum Value [th USD]	2,401,782.90	81,087.58	79,062.26	CTGAN
	Coefficient of Variation [%]	751.81	342.75	261.81	CTGAN
oth_cur_as	Mean [th USD]	688.27	620.79	411.29	CTGAN
	Minimum Value [th USD]	-2964.56	-259.81	-288.09	TVAE
	Quartile I [th USD]	11.97	18.94	21.79	CTGAN
	Median [th USD]	58.80	60.54	55.86	CTGAN
	Quartile III [th USD]	263.94	301.83	261.44	TVAE
	Maximum Value [th USD]	716,969.71	29,415.56	27,022.05	CTGAN
	Coefficient of Variation [%]	854.65	371.19	357.54	CTGAN
ncur_as	Mean [th USD]	2514.05	2399.06	1406.75	CTGAN
	Minimum Value [th USD]	-622.34	-261.92	-622.34	TVAE
	Quartile I [th USD]	0.27	-3.84	7.25	CTGAN
	Median [th USD]	30.60	67.06	67.53	CTGAN
	Quartile III [th USD]	439.81	696.53	407.85	TVAE
	Maximum Value [th USD]	3,543,097.80	115,249.58	115,937.28	TVAE
	Coefficient of Variation [%]	879.50	434.96	442.88	TVAE
tot_as	Mean [th USD]	4633.40	4232.31	2877.63	CTGAN
	Minimum Value [th USD]	-55.34	-55.34	-55.34	BOTH
	Quartile I [th USD]	75.56	52.80	103.29	CTGAN
	Median [th USD]	379.15	290.61	482.89	CTGAN

	Quartile III [th USD]	1675.52	1763.45	1901.62	CTGAN
	Maximum Value [th USD]	5,944,880.70	161,039.44	151,068.43	CTGAN
	Coefficient of Variation [%]	728.97	359.04	274.70	CTGAN
Capital	Mean [th USD]	711.62	499.50	232.78	CTGAN
	Minimum Value [th USD]	-297,119.78	-2269.44	-2019.63	CTGAN
	Quartile I [th USD]	1.33	3.15	-1.36	CTGAN
	Median [th USD]	13.30	26.22	17.63	TVAE
	Quartile III [th USD]	54.54	78.25	36.91	TVAE
	Maximum Value [th USD]	1,235,199.00	53,668.28	45,571.10	CTGAN
	Coefficient of Variation [%]	1317.52	566.85	685.64	TVAE
sh_funds	Mean [th USD]	1917.58	2290.77	1125.07	CTGAN
	Minimum Value [th USD]	-680,837.85	-32,739.08	-47,909.47	TVAE
	Quartile I [th USD]	7.18	-1.44	8.84	TVAE
	Median [th USD]	98.71	114.25	97.77	TVAE
	Quartile III [th USD]	616.49	858.74	547.08	TVAE
	Maximum Value [th USD]	1,814,005.85	117,020.46	110,867.04	CTGAN
	Coefficient of Variation [%]	839.47	383.77	403.90	TVAE
ot_sh_funds	Mean [th USD]	1205.97	906.49	916.94	TVAE
	Minimum Value [th USD]	-924,335.14	-37,356.44	-63,085.24	TVAE
	Quartile I [th USD]	-1.33	-59.31	-8.70	TVAE
	Median [th USD]	55.61	31.24	88.54	CTGAN
	Quartile III [th USD]	427.84	258.86	494.59	TVAE
	Maximum Value [th USD]	1,799,374.78	82,810.24	102,430.68	TVAE
	Coefficient of Variation [%]	1094.36	438.17	430.96	CTGAN
tot_sh_funds	Mean [th USD]	4634.08	4729.18	2857.31	CTGAN
	Minimum Value [th USD]	-350.68	-350.68	-350.68	BOTH
	Quartile I [th USD]	75.56	121.88	94.14	TVAE
	Median [th USD]	378.88	327.45	313.37	CTGAN
	Quartile III [th USD]	1675.45	1643.34	1873.08	CTGAN
	Maximum Value [th USD]	5,944,880.70	174,549.11	153,965.07	CTGAN
	Coefficient of Variation [%]	728.90	372.73	288.93	CTGAN
Stock	Mean [th USD]	590.58	443.74	296.65	CTGAN
	Minimum Value [th USD]	-716.53	-100.44	-93.44	CTGAN
	Quartile I [th USD]	0.00	0.09	-5.39	CTGAN
	Median [th USD]	1.06	14.82	9.29	TVAE
	Quartile III [th USD]	93.12	153.35	25.64	CTGAN
	Maximum Value [th USD]	1,182,202.25	28,632.60	31,595.40	TVAE
	Coefficient of Variation [%]	1106.75	390.92	522.85	TVAE
Debtors	Mean [th USD]	840.52	678.29	545.85	CTGAN

	Minimum Value [th USD]	-295.07	-295.07	-295.07	BOTH
	Quartile I [th USD]	11.17	19.36	24.11	CTGAN
	Median [th USD]	66.78	63.83	69.85	CTGAN
	Quartile III [th USD]	321.15	303.04	318.30	TVAE
	Maximum Value [th USD]	1,182,188.15	43,295.31	37,973.78	CTGAN
	Coefficient of Variation [%]	961.56	428.47	367.82	CTGAN

Source: own elaboration.

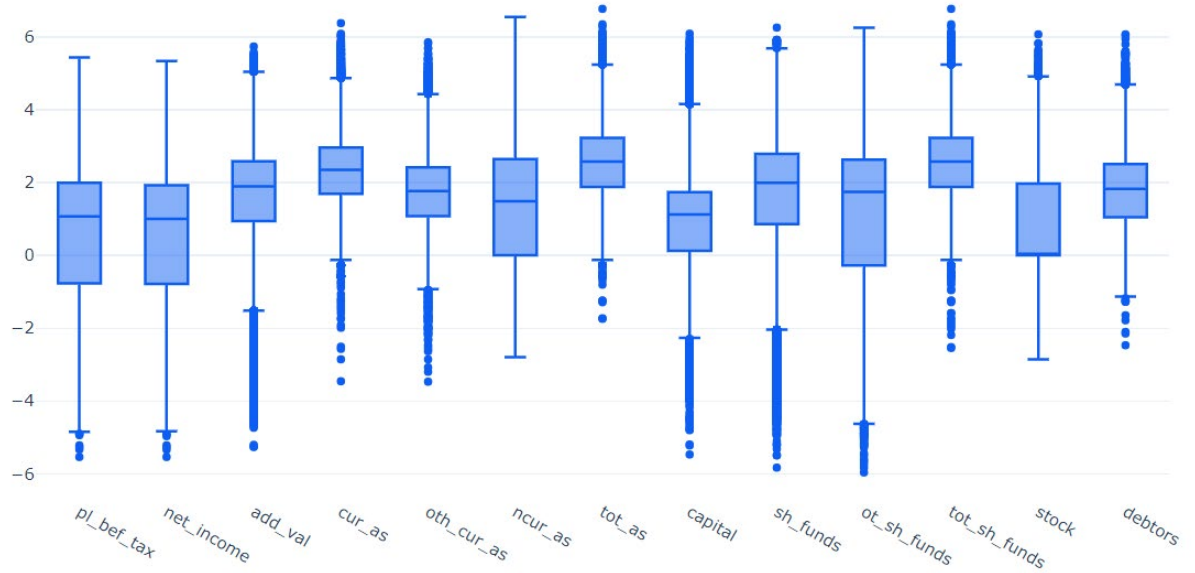

Fig. 10. Boxplots of variables after modified logarithmic transformation from real dataset

Source: own elaboration using Python.

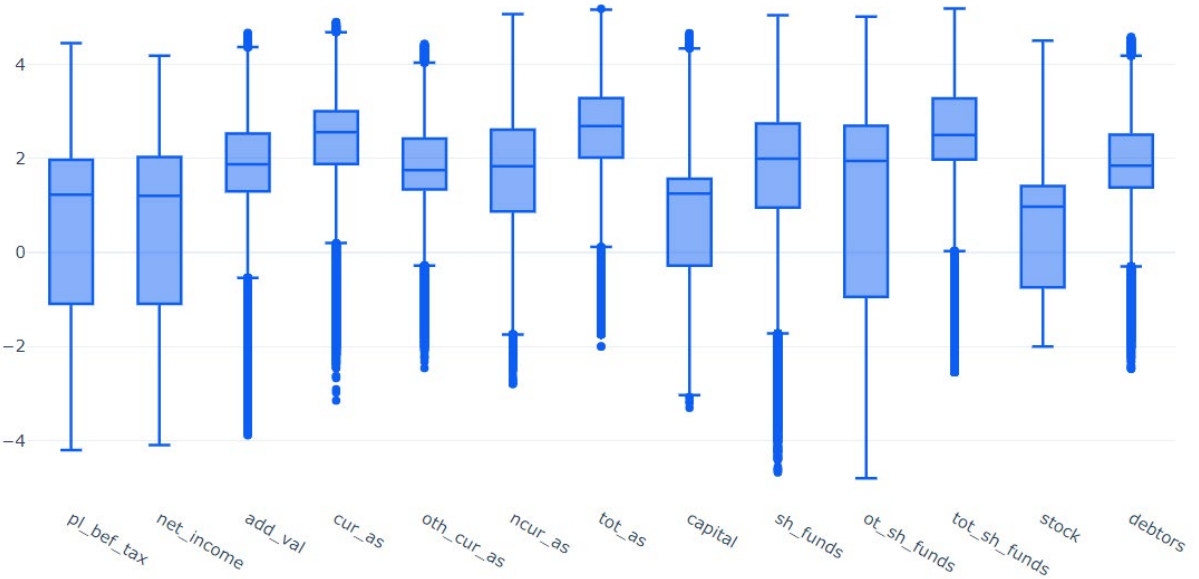

Fig. 11. Boxplots of variables after modified logarithmic transformation from synthetic data generated by TVAE

Source: own elaboration using Python.

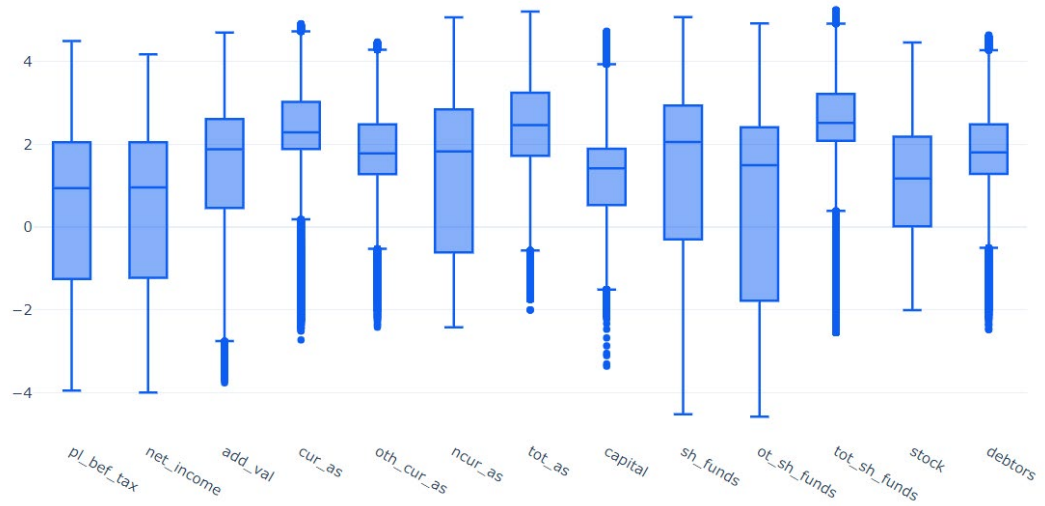

Fig. 12. Boxplots of variables after modified logarithmic transformation from synthetic data generated by CTGAN

Source: own elaboration using Python.

4.4. Principal Component Analysis

A very important reason for generating new datasets was their usefulness in machine-learning algorithms. This allows the use of synthetic data instead of real data in many processes, especially in areas where data protection is crucial. To check it, it was decided to use one of the unsupervised learning methods – principal component analysis. Before applying the PCA method, all the variables were standardised in order to unify their ranges. Table 5 presents the explained variance ratio values for three components for all datasets.

Table 5. The explained variance ratio values depending on the number of principal components

Component	Real data	CTGAN	TVAE
PC1	0.5253	0.5754	0.6096
PC2	0.1947	0.0778	0.0984
PC3	0.0823	0.0576	0.0744
PC1 + PC2	0.7200	0.6532	0.7080
PC1 + PC2 + PC3	0.8023	0.7108	0.7824

Source: own elaboration.

Regarding the sum of the explained variance ratio values for the first two components (PC1 and PC2), the PCA results for the synthetic data generated by the TVAE method gave a similar result to the PCA results for the real data. The value for the first three components (PC1 and PC2 and PC3) also gave better results for the TVAE method.

5. Discussion and Conclusions

To sum up, the concept of synthetic data allows some of the data restrictions such as data privacy, data scarcity and data quality to be overcome. Publishing synthetic data, for example on platforms such as the Kaggle platform (*Kaggle*, n.d.) or hackathons, allows many researchers, students and experts to work at the same time on solutions to cope with a certain problem. This approach, namely open data, helps companies to find multiple solutions while reducing costs and keeping data privacy. For this reason, growing competition will enable providing better final products. Thanks to this, they

will build and test new products (e.g. applications) without access to original, and very often sensitive, data. Synthetic data cannot be published for everyone, but only for one specific academic team or company to solve a problem as a part of a project collaboration, and moreover, it is possible to use them to train advanced models using cloud services, which allows companies and institutions to deal with hardware resources and data privacy restrictions.

Despite the fact that the value of the Overall Quality Score was higher for the data generated by TVAE, when looking in detail at the evaluation metrics, the data generated by CTGAN had a higher quality. In this case, generative adversarial networks turned out to be better than variational autoencoders in terms of statistical properties. Looking at the results of the PCA method for TVAE, the values of the explained variance ratio for the first two and three components were closer for the real data than CTGAN. It is not possible to clearly assess which method is better for generating tabular data as it depends on the situation. For example, to augment the data for breast cancer detection and prognosis, the synthetic data generated by TVAE was slightly better than by CTGAN (Inan et al., 2023).

It is important to obtain a dataset that reflects the real data in many aspects, but it must be remembered that this is not a replica of real data. Unfortunately, in this case both methods (CTGAN and TVAE) failed to deal with the occurrence of really extreme values in the real data. It is worth pointing out that in some analysis, such as anomaly detection, outliers are more important than regular data points. In the process of training models and generating data it is also necessary to take care of minimising overfitting. This can generate records very similar to the real ones, so the reidentification risk increases. Unfortunately, models did not cope automatically with the accepted principles, financial or mathematical. The model itself did not recognise mathematical relations between columns, e.g. in this analysis the sum of all asset-related categories must equal the total assets variable. It is worth noting that the methods used were fairly good at reproducing the distribution of the variables. This is only possible if the input real data copies the proper data distribution (Karbhari et al., 2021). The quality of the synthetic data is strongly related to the quality of the supplied real data.

Generating synthetic data using several methods and then evaluating each new datasets and finally selecting the best for its application is recommended. Such an action allows a solution to be delivered with a high quality. In addition, carrying out the test using multiple values of model parameters enables better results to be obtained. It is proposed that further research should try to use other methods to generate new tabular data, such as CopulaGAN or TableGAN. In this research, datasets without missing values were used, which is rare the real world. It is recommended that the next step is to deal with this preprocessing problem. Perhaps the solution is to carry out a two-stage analysis: first applying data imputation, and then, based on the resulting complete dataset, generating new data. In the future it is also proposed to check the synthetic data usefulness in machine learning models, but using supervised methods, e.g. to solve classification or regression problems. The variables created in this research can be used to create financial indicators, which can later be applied to the construction of company assessment models.

It is also necessary to try to deal with the problem of the mathematical relation between variables. The solution for now may be to generate only these variables that do not depend on others and then calculate new variables based on these. In this article, a solution was proposed for generating numeric data. The real datasets also contain text data, therefore in the next stages of the research it will be very important to deal with this kind of data as well.

References

Awad, M., & Khanna, R. (2015). *Efficient Learning Machines. Theories, Concepts, and Applications for Engineers and System Designers*. Apress Open.

Bamoriya, P., Siddhad, G., Kaur, H., Khanna, P., & Ojha, A. (2022). DSB-GAN: Generation of Deep Learning Based Synthetic Biometric Data. *Displays*, *74*, 102267. https://doi.org/10.1016/j.displa.2022.102267

Bourou, S., El Saer, A., Velivassaki, T.-H., Voulkidis, A., & Zahariadis, T. (2021). A Review of Tabular Data Synthesis Using GANs on an IDS Dataset. *Information*, *12*(9), 375. https://doi.org/10.3390/info12090375

Carvajal-Patiño, D., & Ramos-Pollán, R. (2022). Synthetic Data Generation with Deep Generative Models to Enhance Predictive Tasks in Trading Strategies. *Research in International Business and Finance*, *62*(52), 101747. https://doi.org/10.1016/j.ribaf.2022.101747

Centrum Nowych Technologii dla Polityk Publicznych NASK-PIB. (2022). *Analiza rozwiązań w zakresie anonimizacji danych*. https://www.nask.pl/pl/raporty/raporty/5110,Analiza-rozwiazan-w-zakresie-anonimizacji-danych-i-generowania-danych-syntetyczn.html

Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W. F., & Sun, J. (2017). *Generating Multi-label Discrete Patient Records Using Generative Adversarial Networks*, *68*, 1-20. http://arxiv.org/abs/1703.06490

Efimov, D., Xu, D., Kong, L., Nefedov, A., & Anandakrishnan, A. (2020). *Using Generative Adversarial Networks to Synthesize Artificial Financial Datasets*. *NeurIPS*, 3-10. http://arxiv.org/abs/2002.02271

European Parliament. (2023, August 6). *EU AI Act: First Regulation on Artificial Intelligence*. www.europarl.europa.eu. Retrieved November 4, 2023 from https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

Figueira, A., & Vaz, B. (2022). Survey on Synthetic Data Generation, Evaluation Methods and GANs. *Mathematics*, *10*(15), 1-41. https://doi.org/10.3390/math10152733

Hameed, M. A. B., & Alamgir, Z. (2022). Improving Mortality Prediction in Acute Pancreatitis by Machine Learning and Data Augmentation. *Computers in Biology and Medicine*, *150*, 106077. https://doi.org/10.1016/j.compbiomed.2022.106077

Inan, M. S. K., Hossain, S., & Uddin, M. N. (2023). Data Augmentation Guided Breast Cancer Diagnosis and Prognosis Using an Integrated Deep-Generative Framework Based on Breast Tumor's Morphological Information. *Informatics in Medicine Unlocked*, *37*, 101171. https://doi.org/10.1016/j.imu.2023.101171

Kaggle: Your Machine Learning and Data Science Community. (n.d.). Retrieved October 28, 2023 from https://www.kaggle.com/

Karbhari, Y., Basu, A., Geem, Z. W., Han, G.-T., & Sarkar, R. (2021). Generation of Synthetic Chest X-ray Images and Detection of COVID-19: A Deep Learning Based Approach. *Diagnostics*, *11*(5), 895. https://doi.org/10.3390/diagnostics11050895

Khaemba, N., Traoré, I., & Mamun, M. (2023). A Framework for Synthetic Agetech Attack Data Generation. *Journal of Cybersecurity and Privacy*, *3*(4), 744-757. https://doi.org/10.3390/jcp3040033

Kingma, D. P., & Welling, M. (2013). *Auto-Encoding Variational Bayes* (2nd International Conference on Learning Representations, ICLR 2014 – Conference Track Proceedings, Ml, pp. 1-14). http://arxiv.org/abs/1312.6114

Li, J., Wang, X., Lin, Y., Sinha, A., & Wellman, M. P. (2020). *Generating Realistic Stock Market Order Streams* (AAAI 2020 – 34th AAAI Conference on Artificial Intelligence, pp. 727-734). https://doi.org/10.1609/aaai.v34i01.5415

Liu, C. (2022). Risk Prediction of Digital Transformation of Manufacturing Supply Chain Based on Principal Component Analysis and Backpropagation Artificial Neural Network. *Alexandria Engineering Journal*, *61*(1), 775-784. https://doi.org/10.1016/j.aej.2021.06.010

Lu, Y., Shen, M., Wang, H., van Rechem, C., & Wei, W. (2023). *Machine Learning for Synthetic Data Generation: A Review*. *14*(8), 1-19. http://arxiv.org/abs/2302.04062

Muñoz-Cancino, R., Bravo, C., Ríos, S. A., & Graña, M. (2022). Assessment of Creditworthiness Models Privacy-Preserving Training with Synthetic Data. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 13469 LNAI* (pp. 375-384). https://doi.org/10.1007/978-3-031-15471-3_32

Murtaza, H., Ahmed, M., Khan, N. F., Murtaza, G., Zafar, S., & Bano, A. (2023). Synthetic Data Generation: State of the Art in the HealthCare Domain. *Computer Science Review*, *48*, 100546. https://doi.org/10.1016/j.cosrev.2023.100546

Neves, D. T., Alves, J., Naik, M. G., Proença, A. J., & Prasser, F. (2022). From Missing Data Imputation to Data Generation. *Journal of Computational Science*, *61*, 101640. https://doi.org/10.1016/j.jocs.2022.101640

Pathare, A., Mangrulkar, R., Suvarna, K., Parekh, A., Thakur, G., & Gawade, A. (2023). Comparison of Tabular Synthetic Data Generation Techniques Using Propensity and Cluster Log Metric. *International Journal of Information Management Data Insights*, *3*(2), 100177. https://doi.org/10.1016/j.jjimei.2023.100177

Patki, N., Wedge, R., & Veeramachaneni, K. (2016). The Synthetic Data Vault. *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 399-410. https://doi.org/10.1109/DSAA.2016.49

Podolszańska, J. (2021). *Proces rekonstrukcji obrazu tomograficznego w oparciu o sieć Variational Autoencoder*, *10*(1), 61-64.

Rajotte, J.-F., Bergen, R., Buckeridge, D. L., El Emam, K., Ng, R., & Strome, E. (2022). Synthetic Data as an Enabler for Machine Learning Applications in Medicine. *IScience*, *25*(11), 105331. https://doi.org/10.1016/j.isci.2022.105331

Shahbazian, R., & Trubitsyna, I. (2022). DEGAIN: Generative-Adversarial-Network-Based Missing Data Imputation. *Information*, *13*(12). https://doi.org/10.3390/info13120575

Sivakumar, J., Ramamurthy, K., Radhakrishnan, M., & Won, D. (2023). GenerativeMTD: A Deep Synthetic Data Generation Framework for Small Datasets. *Knowledge-Based Systems*, *280*, 110956. https://doi.org/10.1016/j.knosys.2023.110956

Statista. (2023, August 22). *Data Created Worldwide 2010-2025*. Statista. Retrieved October 20, 2023 from https://www.statista.com/statistics/871513/worldwide-data-created/

Synthetic Data Metrics. Version 0.11.1. DataCebo, Inc. https://docs.sdv.dev/sdmetrics/

Synthetic Data Vault. Version 1.5.0. DataCebo, Inc. https://docs.sdv.dev/sdv/

Szymura, A. (2022). Risk Assessment of Polish Joint Stock Companies: Prediction of Penalties or Compensation Payments. *Risks, 10*(5). https://doi.org/10.3390/risks10050102

Talabis, M. R. M., McPherson, R., Miyamoto, I., Martin, J. L., & Kaye, D. (2015). Analytics Defined. *Information Security Analytics*, 1-12. https://doi.org/10.1016/b978-0-12-800207-0.00001-0

The European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Da. Official Journal of the European Union (OJ L 119 4.5.2016, p. 1).

Wang, Y., Li, D., Li, X., & Yang, M. (2021). PC-GAIN: Pseudo-label Conditional Generative Adversarial Imputation Networks for Incomplete Data. *Neural Networks*, *141*, 395-403. https://doi.org/10.1016/j.neunet.2021.05.033

Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling Tabular Data Using Conditional GAN. *Advances in Neural Information Processing Systems*, *32*(NeurIPS). http://arxiv.org/abs/1907.00503

Yilmaz, B., & Korn, R. (2022). Synthetic Demand Data Generation for Individual Electricity Consumers: Generative Adversarial Networks (GANs). *Energy and AI*, *9*, 100161. https://doi.org/10.1016/j.egyai.2022.100161

Yoon, J., Jordon, J., & van der Schaar, M. (2018). *GAIN: Missing Data Imputation using Generative Adversarial Nets* (35th International Conference on Machine Learning, ICML 2018, 13, 9042-9051). http://arxiv.org/abs/1806.02920

Syntetyczne dane finansowe: studium przypadku dla danych polskich spółek z ograniczoną odpowiedzialnością

Streszczenie

Cel: Celem artykułu jest prezentacja i ocena koncepcji danych syntetycznych. Są to całkowicie nowe, sztucznie wygenerowane dane, ale zachowujące własności statystyczne danych rzeczywistych. Ze względu na ich statystyczne podobieństwo do danych rzeczywistych mogą być wykorzystywane zamiast nich. Pozwala to na udostępnianie danych na zewnątrz z jednoczesnym zagwarantowaniem ich prywatności.

Metodyka: Nowe zbiory wygenerowano na bazie informacji finansowych polskich spółek z ograniczoną odpowiedzialnością. Wszystkie potrzebne dane wejściowe pochodzą z bazy Orbis i dotyczą 2020 roku. Do tworzenia danych syntetycznych zdecydowano się wykorzystać modele generatywne: CTGAN (oparte na architekturze GAN) i TVAE (oparte na autoenkoderach). Finalnie porównano otrzymane dane syntetyczne z rzeczywistymi pod kątem własności statystycznych (np. podobieństwo rozkładów, korelacje) oraz ich możliwości zastosowania w analizie danych (PCA).

Wyniki: Ogólny wskaźnik oceny jakości danych był wyższy dla danych wygenerowanych metodą TVAE, ale zagłębiając się w szczegóły, stwierdzono, że dane wygenerowane metodą CTGAN są lepszej jakości pod względem zachowania własności statystycznych w stosunku do danych rzeczywistych. Po porównaniu wyników metody PCA ponownie stwierdzono, że TVAE okazało się lepsze niż CTGAN. Dodatkowo metoda TVAE była mniej czasochłonna niż CTGAN.

Implikacje i rekomendacje: Przed udostępnieniem danych syntetycznych na zewnątrz zaleca się wygenerowanie ich z wykorzystaniem kilku algorytmów, porównanie ich wyników końcowych, a następnie – na ich podstawie – wybranie jednej, najlepszej opcji. Takie działanie pozwoli na otrzymanie zbioru o najwyższej jakości. W przyszłych badaniach proponuje się sprawdzenie innych algorytmów (np. CopulaGAN lub TableGAN) oraz podjęcie próby poradzenia sobie z rzeczywistymi problemami występującymi w danych, które zostały pominięte w tej analizie, jak np. występowanie braków danych (w tym artykule pracowano na kompletnym zbiorze danych). Dane wygenerowane w tym badaniu mogą być wykorzystane do budowy wskaźników finansowych, które z kolei mogą być później zastosowane w tworzeniu modeli oceny przedsiębiorstw.

Oryginalność/wartość: Dane syntetyczne pomagają przezwyciężyć liczne ograniczenia, jak np. prywatność danych czy ich niedobór. Ze względu na ich statystyczne podobieństwo do danych rzeczywistych możliwe jest użycie ich w zaawansowanych modelach uczenia maszynowego zamiast danych rzeczywistych. Analiza na dobrych jakościowo danych syntetycznych pozwala na osiągnięcie podobnych wniosków co analiza przeprowadzana na danych rzeczywistych, z zachowaniem przy tym prywatności danych, bez udostępniania danych wrażliwych osobom trzecim.

Słowa kluczowe: dane syntetyczne, modele generatywne, dane finansowe, CTGAN, TVAE