Irving Fisher Committee on
Central Bank Statistics

BIS

IFC-Bank of Italy Workshop on "Data science in central banking"

18-20 February 2025

Generalised weighted framework for synthetic data evaluation[1]

Chiung Ching Ho,
Central Bank of Malaysia

[1] This contribution was prepared for the workshop. The views expressed in this publication are those of the authors and do not necessarily represent the official views of the Committee, its members, or the BIS.

Generalised weighted framework for synthetic data evaluation

Chiung Ching Ho (BNM)

The views expressed in this paper are of the author and do not necessarily represent the views of the associated institutions.

Abstract

The application of synthetic data in central banking has been on the rise, particularly as a means of data sharing securely sharing data between central banks and governmental agencies or research institutions. The need to securely share data is driven by increasing legislation worldwide concerning data protection and privacy. Another emerging reason for using synthetic data in central banks is to increase data availability, in situations where the original data is scarce or difficult to collect

As synthetic data sharing increases, it becomes imperative that the correctness and quality of the synthetic is assessed by data providers prior to publication. In this paper, a generalised weighted framework for synthetic data evaluation is proposed to enable this objective. This framework aims to combine measures for four-dimensions:

i. Statistical distance
ii. Structural similarity
iii. Machine learning task performance
iv. Privacy

The composite final weighted score = wSD + wSS +wML + wP , where SD denotes statistical distance, SS denotes structural similarities, ML denotes machine learning task performance and P denotes privacy, with w being the corresponding weight for each dimension. Recent works for synthetic data evaluation has focused on the statistical distance of the synthetic data from the original data and its structure, which intuitively impacts machine learning outcomes or privacy concerns but not all these dimensions together.

This framework was evaluated on a credit-card default dataset with identifying information and content that is suitable for classification, clustering, and time-series forecasting. A synthetic dataset was created from the original dataset, using a variety of methods including statistical and deep learning methods, as well as differential privacy methods. Subsequently, the synthetic datasets were compared to the original dataset using the four dimensions mentioned above during experiments for classification, clustering and time-series forecast. Ablative experiments were further conducted to evaluate the framework's ability to cater to different scenarios.

In the ablative experiments, the framework was adjusted for different scenarios e.g. different prioritisation for different analytical outcomes, or privacy-first analysis ahead of analytics. This generalised approach will allow users to conduct differentiated evaluation on synthetic data. This in-turn will help data providers

provide high-quality synthetic data, which is well suited to the task on hand, leading to efficiency gains.

JEL classification: C18, C45, C63

Contents

1. Introduction

Data collected by central banks serve as essential resources that have gained considerable value with the increasing application of artificial intelligence (AI). This data is especially valuable because it is typically not published as open data, due to data governance policies and legislation (European Union, 2024, p. 10) designed to protect privacy and confidentiality of data. Consequently, this data can be utilized by central banks to enhance existing large language models (LLMs), which are experiencing a scarcity of data (Robinson, 2024).

Simultaneously, data sharing conducted by central banks remains influenced by global events such as the Great Financial Crisis (GFC) of 2007-2009 and, more recently, the Covid-19 pandemic. Following the GFC, seven key recommendations have been

identified to enhance data-sharing efforts (Irving Fischer Committee on Central Bank Statistics, 2023):

1. Promoting the use of common statistical identifiers.
2. Promoting the exchange of experience on statistical work with granular data and improving transparency
3. Balancing confidentiality and users' needs.
4. Linking different datasets.
5. Provision of data at the international level.
6. Consideration of ways of improved data-sharing of granular data; and
7. Collection of data only once.

To balance the need for confidentiality with users' requirements, central banks have adopted the practice of sharing data at an aggregated level and providing granular data in a depersonalised format. While the former method maintains confidentiality and privacy, it is less effective than granular data for conducting timely and thorough risk analysis, as well as formulating targeted policy measures (Ling, 2013).

Recent developments have seen the usage of synthetic data as a solution for sharing data whilst adhering to privacy and confidentiality concerns. Synthetic data is data that has been generated using a purpose-built mathematical model or algorithm, with the aim of solving a (set of) data-science task(s) (Jordon et al., 2022). Here, synthetic data is contrasted with real data. Real data is data generated not by a model but by a real-world system, e.g. data that is submitted by a central bank's regulatees.

As synthetic data is statistically or semantically similar to the originally data used in its creation, an argument is made that the synthetic data will then be able to be used in the same way for machine learning applications, and with reasonable expectations that the outcome would be similar to outcomes using the original data. At the same time, the synthetic data set would be able to preserve privacy and confidentiality. As central banks have a mission to manage reputational risk, it is imperative that any data synthetic data set is assessed for utility, fidelity and privacy prior to being shared.

In this paper, a framework is proposed to address this problem and apply it on an open-sourced credit card dataset that has been widely used for machine learning experiments. The framework proposes a way to assess synthetic data sets from the perspective of utility, fidelity and privacy. Utility is measured as fitness-of-purpose for the synthetic data to be used for machine learning applications, while fidelity is measured as the statistical distance and structural similarities between the synthetic data and the original data. Finally, privacy is measured as the statistical distance between the synthetic data and the original data. This combination fulfils the conflicting requirements for synthetic data to be meaningful, in that it both sufficiently similar and different from the original data. This framework also proposes a weighted average of the different measures for utility, fidelity and privacy for greater flexibility. In terms of originality, as far as it is known, this framework is the first framework which introduces a weighting mechanism to fine-tune the assessment of the synthetic dataset for privacy, utility or fidelity.

2. Synthetic data

Synthetic data is data that has been generated using a purpose-built mathematical model or algorithm, with the aim of solving a (set of) data-science task(s). Here, synthetic data is contrasted with real data. Real data is data generated not by a model but by a real-world system, e.g. data that is submitted by a central bank's regulatees.

The synthetic data generator may include models such as gaussian mixed models (GMM), Variational Auto-encoders (VAE), Copula models, agent based or econometric models or even stochastic differential equations. Models which are used in this paper are mentioned in Section 3.2.

Synthetic data has proven to be valuable in scenarios where real data is limited, for instance, when generating synthetic data derived from original datasets that were costly to obtain or within financial systems that have only recently commenced data collection. In such cases, the synthetic data serves as training data for machine learning applications through a process known as data augmentation. Effective data augmentation methodologies must also address the elimination or reduction of biases that may arise due to historical biases, while maintaining statistical accuracy.

This paper will focus on the utilization of synthetic data for data sharing purposes. Machine learning solutions fundamentally rely on high-quality training data, including data owned by central banks. In this context, it would be beneficial to share synthetic versions of the relevant data. It is essential for the synthetic data to exhibit statistical similarity to the original data, as the underlying assumption is that machine learning models trained on statistically similar synthetic data will yield results comparable to those obtained using the original data.

2.1 Synthetic data and central bank

In this section, the literature concerning the usage of synthetic data by central bank practitioners is presented. From the literature search, it is observed that current efforts on using synthetic data in central banks has focused on either enabling access to data necessary for research or has focused on assessment on synthetic data from a utility perspective.

The Deutsche Bundesbank Research Data and Service Centre (RDSC) provides researchers with access to micro data for research purposes, up to individual data with the corresponding restrictions. To overcome further confidential restrictions as well as to enable data sharing to be done with researchers who are unable to access the, synthetic data of micro data such as the Survey of Income and Program Participation is created by adding noise to the data (Drechsler, 2011)

Banking microdata collected by the Central Bank of Paraguay from 2017-2023 in the areas of financial inclusion, granular term deposits, and credit risk management were used to create synthetic data for the purpose of data sharing. The synthetic data created were evaluated for utility in these areas (Caceres & Moews, 2024).

Banco de Espana has also conducted preliminary efforts to create synthetic data and assessed the synthetic data for both utility and confidentiality. The targeted data was microdata for individual non-financial firms and microdata for loans to legal persons (Lapteva, 2023).

Bangko Sentral ng Pilipinas (Ong et al., 2024) has also reported their efforts to create synthetic data for the purposes of data sharing. In their work, the Consumer Expectation Survey was chosen to be used as the data to create synthetic data. In their work, assessment was performed on the synthetic data from a fidelity, utility and privacy perspective.

From the literature, central banks have shown increasing appetite to use synthetic data as a means of sharing data. Early efforts have focused on micro-data, with an interest on utility and privacy rather than fidelity.

2.2 Models for synthetic data generation

The models for creating synthetic data or synthetic data generators (SDG) are varied. It ranges from rule-based systems, random sampling, statistical approaches, differential privacy approaches, and machine learning approaches. A comprehensive survey of all these methods are presented in (Goyal & Mahmoud, 2024).

Of the machine learning approaches, generative adversarial networks (GANs), variational autoencoders (VAEs), and large language models (LLMs) are the most popular in recent years.

Generative Adversarial Networks (GANs) provide a powerful method for generating synthetic data using machine learning. They consist of two components: a generator that creates data and a discriminator that evaluates its authenticity. Through a process of iterative competition, the generator improves its ability to produce data that closely mimics real-world datasets.

In contrast, Variational Autoencoders (VAEs) take a probabilistic approach to data generation. They encode input data into a compressed latent space and then decode it back, enabling the creation of new data points sampled from this latent space. This ensures that the synthetic data retains the key patterns and variability found in the original dataset.

Large language models (LLMs) like generative pre-trained transformer (GPT) offer another approach, particularly for text-based data. By training on vast datasets, these models can generate text that is coherent and contextually accurate, often indistinguishable from text written by humans. This makes them highly effective for applications requiring natural language processing.

From the papers published by central banks, SDG that were used included gaussian diffusion models (GDM) , synthetic minority oversampling technique (SMOTE), gaussian Copula (GC), conditional tabular generative adversarial network (CTGAN), tabular variational autoencoder (TVAE) and gaussian mixture models (GMM).

2.3 Assessing synthetic data

Any synthetic data that is generated using an SDG needs to be assessed. This is more so for synthetic data that is produced and shared by central banks, due to the need for central banks to maintain the highest levels of trust to execute their mandates effectively. Performance metrices for synthetic data in the areas of utility, privacy, utility and fidelity are described in (Osorio-Marulanda et al., 2024).

2.3.1 Privacy

Privacy metrics for synthetic data aim to quantify the potential disclosure risk regarding the original source data. The required stringency and type of metric depend heavily on the context; metrics suitable for internal use may differ from those needed for public releases.

Primarily, robust privacy metrics are derived from frameworks like differential privacy. These metrics typically evaluate the data generation algorithm, providing a mathematically rigorous measure (e.g., privacy budget parameters like ϵ and δ) of the maximum statistical information leakage allowed by the process across all outputs. They offer a systematic way to analyse the inherent privacy safeguards of the method used.

A significant limitation, however, exists in developing reliable metrics to assess the privacy of a specific, already generated synthetic dataset, especially without knowledge of its provenance. Current state-of-the-art metrics characterise the properties of the generation process itself, rather than providing a definitive privacy score for a single data instance.

It is with this in mind that the work in this paper used the T-measure, as it measures the outcome of the generative process.

2.3.2 Utility

Utility evaluation metrics encompass techniques that evaluate the general performance of models, algorithms, or data. It involves evaluating how well the utility of the original data is maintained in the synthetic version of the data. Traditional metrices related to machine learning operations such as accuracy, precision, recall, F1 score, area under curve (AUC) are useful utility metrices.

Utility measures which have been reported by central banks for assessing their synthetic data efforts includes accuracy, AUC, recall, precision, F1 scores , Kappa, Matthews Correlation Coefficient (MCC) and error measures such as root mean square error (RMSE). The choice of a utility evaluation measure will be determined by the use case in which the synthetic data is to be used.

The utility evaluation measure is frequently compared to the privacy evaluation measure. It is crucial for balancing the trade-offs between privacy and utility in various applications.

2.3.3 Fidelity

Fidelity refers to metrics that directly compare the synthetic dataset with the real one, rather than through a model or task performance. It assesses how well the synthetic data statistically matches the original data. These measures are often used based on the intuition that a specific level of fidelity will lead to better performance across various machine learning tasks.

3. Methods

3.1 Framework

The purpose of the proposed framework is to perform assessment of the dataset created by the SDG for the dataset's privacy, utility and fidelity.

The framework is divided into three distinct phases, which is described in more detail below:

Phase 1 - Data Preprocessing

In this phase, the raw dataset was ingested for analysis, after which non-informative features like unique identifiers (e.g. ID columns) were systematically removed to prevent data leakage. The remaining features were then divided into two categories: categorical features (e.g., pay status) and continuous features (e.g., age). Following this, feature-specific preprocessing techniques were applied; categorical variables underwent transformation via one-hot encoding to enable appropriate interpretation by machine learning models, while continuous variables were normalised using Min-Max scaling, rescaling values to a [0,1] range for uniformity and stability during model training. This preprocessing phase resulted in a fully cleaned and standardized dataset suitable for the subsequent phase.

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Figure X: Min-Max scaling

Where:

- X is the original value,
- X_{min} is the minimum value in the dataset,
- X_{max} is the maximum value,
- $X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}}$ is the normalised value in the range [0,1][0, 1][0,1].

Phase 2 - Synthetic Data Generation

The pre-processed dataset from Phase 1 is then utilized as the basis for synthetic data generation, beginning with the extraction of feature vectors to serve as input for the SDG . The SDG is fitted to the feature distribution to learn the underlying probability density of the original data. Subsequently, new synthetic samples were generated by sampling from the trained SDG. To maintain interpretability, these synthetic samples were inversely transformed back to the original feature scales where applicable. Finally, synthetic labels were assigned to the generated data points using a pre-trained classifier, ensuring consistency with the original label distribution. This phase resulted in a synthetic dataset designed to closely mimic the statistical and structural properties of the original dataset

Phase 3 – Evaluation

In this phase, the synthetic dataset created by the SDG was evaluated for privacy, utility and fidelity across four key dimensions:

(i) Statistical distance (for fidelity) was assessed using the Wasserstein distance between real and synthetic data distributions.

The Wasserstein distance is given by the following formula,

$$W(p,q) = \inf_{\gamma \in \Gamma(p,q)} \mathbb{E}_{(x,y)\sim\gamma}[\|x-y\|]$$

Where:

- $p\ and\ q$ are the distributions (often represented as probability measures),
- $\Gamma(p,q)$ is the set of all **joint distributions** (couplings) with marginals $p\ and\ q$
- $[\|x-y\|]$ is typically the **Euclidean distance** between x and y,
- The infimum finds the optimal transport plan (minimum cost to transform one distribution into the other)

For 1D empirical distributions (single column), the Wasserstein-1 distance simplifies to:

$$W(p,q) = \int_0^1 |\, F^{-1}(t) - G^{-1}(t)\,|\, dt$$

Where:

- F^{-1} and G^{-1} are the quantile functions (inverse CDFs) of p and q

The final score is $S_{SD} = 1 - \frac{1}{d}\sum_{i=1}^{d} W_1^{(i)}$ where :

- d is the number of columns (features),
- $W_1^{(i)}$ is the **Wasserstein distance** between the i-th feature in the real and synthetic datasets,
- $S_{SD}\ \ is$ the normalised **Statistical Distance Score**: higher values indicate better similarity.

(ii) Structural similarity (for fidelity) was investigated via comparisons of feature correlations and marginal distributions between the real and synthetic data.

The **correlation similarity** compares Pearson correlation matrices from real ρ_{real} and synthetic ρ_{syn} datasets. The mean absolute error (MAE) between the matrices is used to compute the score:

$$S_{\text{corr}} = \frac{1}{1 + \text{MAE}(\rho_{\text{real}}, \rho_{\text{syn}})}$$

Additionally, **distribution similarity** is assessed using the Kolmogorov-Smirnov (KS) test for each feature. For a given column, the KS statistic D_n is calculated as the maximum difference between empirical CDFs, and the normalised score is:

$$S_{\text{KS}}^{(i)} = 1 - D_n^{(i)}$$

The final normalised structural similarity score combines these two measures:

$$S_{SS} = \frac{1}{2}\left(S_{\text{corr}} + \frac{1}{d}\sum_{i=1}^{d} S_{\text{KS}}^{(i)}\right)$$

(iii) Machine learning utility was then measured through downstream task performance, encompassing classification accuracy, clustering quality, and time series analysis.

For a classification task (in this work, it was classification of the default payment label using a Random Forest classifier), the S_{ML} normalised Machine Learning Score is given by

$$S_{ML} = 0.5 \cdot (\text{Accuracy} + \text{F1})$$

Additional unsupervised tasks, such as clustering (K-Means) and time series forecasting (Linear Regression on lagged financial features), were evaluated using Silhouette Score and Mean Squared Error (MSE), respectively but were not included in the ML score.

(iv) Privacy assessment was conducted using the T-measure score to estimate the risk of re-identification, with the overall T-measure score across all features calculated as the average of the absolute T-measure.
The formula for the T-measure for a single feature is given by

$$T = \frac{\bar{x}_r - \bar{x}_s}{\sqrt{\frac{s_r^2}{n_r} + \frac{s_s^2}{n_s}}}$$

Where:

- $\bar{x}_r$ = mean of the feature in the **real** dataset
- $\bar{x}_s$ = mean of the feature in the **synthetic** dataset
- s_r^2, s_s^2 = variances of the feature in real and synthetic datasets respectively
- n_r, n_s = number of samples in real and synthetic datasets

The formula for the overall T-measure score across all features is given by

$$T_{\text{score}} = \frac{1}{d}\sum_{i=1}^{d} | T_i |$$

Where:
d = number of features
T_i = t-measure for feature i

For each numerical feature, the standardized mean difference is calculated as:

$$T_{\text{col}} = \frac{| \mu_{\text{syn}} - \mu_{\text{real}} |}{\sigma_{\text{real}}}$$

where μ and σ represent mean and standard deviation, respectively.
The overall privacy score is:

$$S_P = \frac{1}{1 + \frac{1}{d}\sum_{i=1}^{d} T_{\text{col}}^{(i)}}$$

Each of these evaluation metrics was assigned an equal weight of 25% to compute a composite final weighted score, providing a balanced assessment of the synthetic data's quality by capturing its ability to preserve essential patterns, support model training, and protect privacy.

$$\text{Composite Final Weighted Score} = \sum_{i \in \{SD, SS, SML, SP\}} w_i \cdot S_i$$

The weightage may be adjusted according to the use case for which the synthetic data , adjusting for the weightage for privacy (privacy score), utility (machine learning score and statistical distance score) and fidelity (structural similarity score).

3.2 Experiment

To demonstrate the usage of the generalised weighted framework for synthetic data evaluation, experiments were conducted to firstly generate the synthetic data set using different SDG, and secondly to use the framework to elicit the weighted score.

SDG strengths and weaknesses

Table 1

Name of SDG	Strengths	Weaknesses
Gaussian Diffusion Models (GDM)	• High flexibility in modelling distributions Good for high-dimensional data	• Computationally intensive Requires substantial amounts of training data
Gaussian Copula (GC)(Patki et al., 2016)	• Maintains statistical properties Suitable for tabular data	• Assumes a Gaussian dependence structure Struggles with complex, non-linear relationships
Conditional Tabular GAN (CTGAN)(Xu et al., 2019)	• Manages imbalanced and multi-modal data Captures complex dependencies	• Training instability Sensitive to hyperparameter tuning
Tabular Variational Autoencoder (TVAE)(Xu et al., 2019)	• Effective in capturing latent structures Suitable for tabular data	• May overfit on small datasets Needs careful parameter tuning
Gaussian Mixtures Models (GMM)(Dempster et al., 1977)	• Model data as a combination of distributions Interpretable and straightforward	• Assumes data can be represented as Gaussian mixtures Sensitive to initialization
Time Dependent Self Attention (TDSA)*	• Strong at time-series data generation Preserves temporal correlations	• Limited to time-series data Computationally expensive for large datasets
Tabular Diffusion Probabilistic Model (TabDPM)(Kotelnikov et al., 2024)	• Robust for tabular data with complex dependencies Flexible with varying distributions	• Computationally heavy Requires careful parameter configuration

Notes: The TDSA algorithm is the author's own algorithm that a neural network with a self-attention mechanism to predict noise added at different time steps based on the noisy data and the current time step. Data generation starts by gradually adding noise to real data over many steps in a forward diffusion process. Finally, new data is synthesized by starting with random noise and iteratively using the neural network's predictions to reverse the noising process step-by-step.

3.2.1 Data set

The synthetic data originates from an open-source credit card dataset chosen for its suitability in assessing utility through machine learning applications. This dataset contains 30,000 entries (excluding the header) with information about credit card clients (Yeh & Lien, 2009). Key features encompass core financial details like credit limit balance, bill amounts for six consecutive months, and payment amounts for the same six months. Additionally, it includes demographic information such as the client's gender, age, education level, and marital status , along with a seven-month payment status history. The target variable is a binary indicator (0 or 1) labelled "Default payment next month," which predicts potential client default in the following month.

3.2.2 Experimental Results

Experiments were conducted to evaluate the proposed framework whereby seven SDG were used to create 30,000 synthetic elements using the original or real dataset, $dataset_{real}$. The synthetic dataset, $dataset_{syn}$ were evaluated using the measures described in Section 3.3.3, to elicit the normalised scores for the statistical distance, structural similarity, machine learning and privacy measures. The parameters of the experiments are as follows

1. Measures for statistical distance, structural similarity, machine learning and privacy measures are equal at 0.25 (Figures 1-4)
2. Time-series and clustering tasks are added to the classification tasks (Figure 5)
3. Measures for privacy and machine learning are over-weighted (Figure 6)

Overall weighted score for various SDG

Figure 1

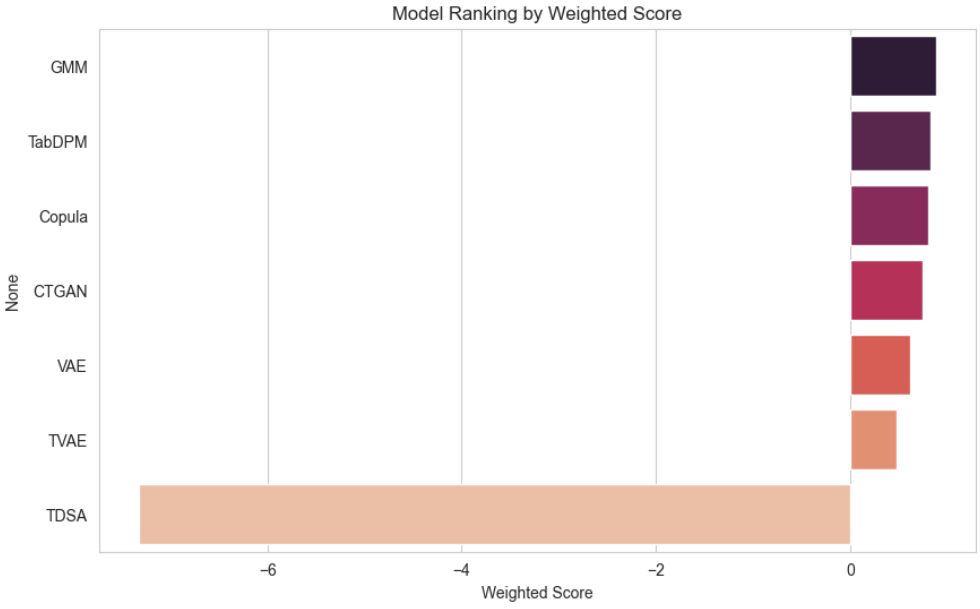

Figure 1 shows that the GMM performed the best, followed by TabDPM and GC. All three models showed a weighted score of above 0.80 on a scale of 0 to 1.0. This shows that the evaluation scheme can measure the privacy, utility and fidelity of the

synthetic dataset using a single measure. The poor results of the synthetic dataset generated by TDSA can be attributed to its high Wasserstein distance score

Privacy scores for various SDG

Figure 2

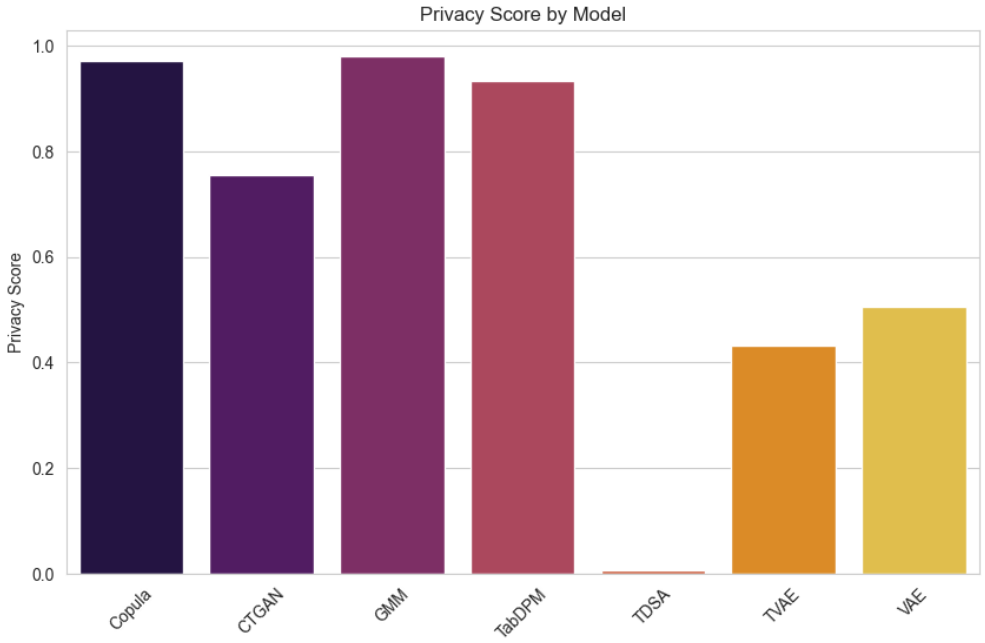

Figure 2 shows that the GC, GMM and TabDPM SDG resulted in synthetic data sets which has high privacy scores exceeding 0.8 on a scale of 0 to 1.0. TDSA which focuses on time-series with self-attention performed the worst

Utility scores for various SDG

Figure 3

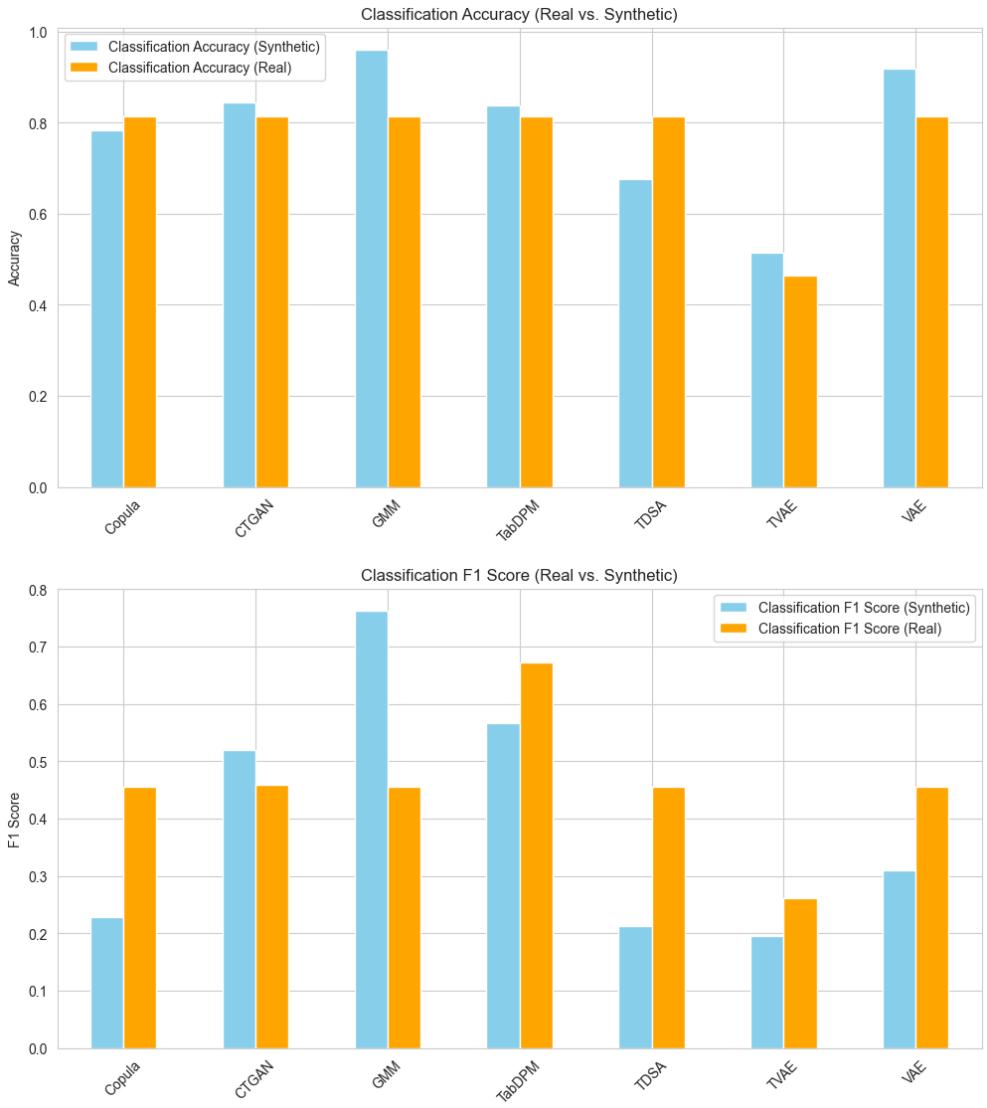

Figure 3 shows the utility scores for various datasets created using the same set of SDGs. For classification measured using accuracy, five out of seven SDG resulted in synthetic datasets which returned higher classification scores as compared to the original dataset. For classification measured using the F-1 score, four out of seven SDG resulted in the real dataset returning higher classification scores.

Fidelity scores for various SDG

Figure 4

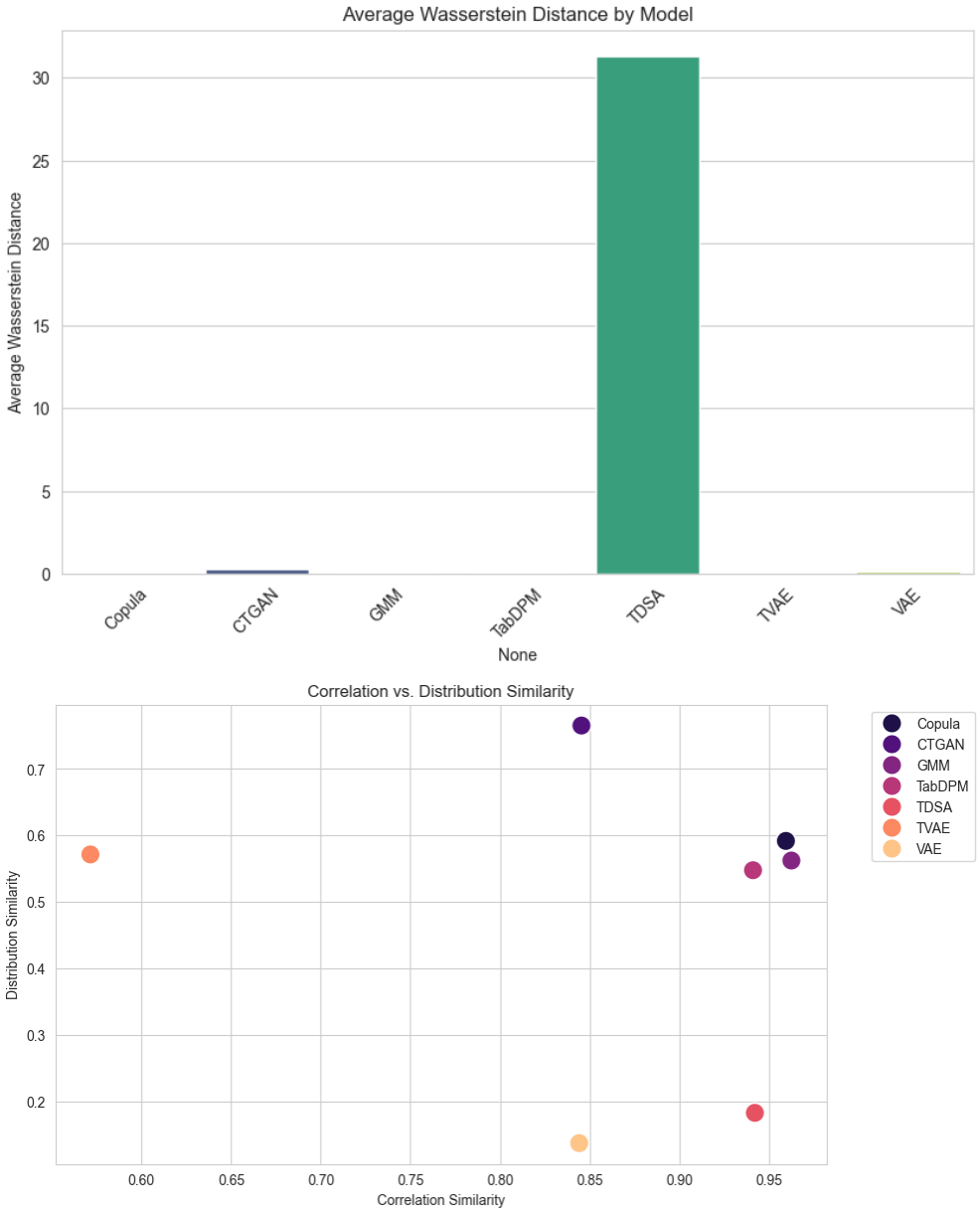

Figure 4 shows the results for synthetic datasets created using the same set of SDGs. Four out of seven SDGs resulted in synthetic datasets which has high correlation similarity (above 0.9 on a scale of 0-1.0), while only CTGAN resulted in a synthetic dataset which has distribution similarity of above 0.7 on a scale of 0 to 1.0.

Privacy scores vs combined ML score

Figure 5

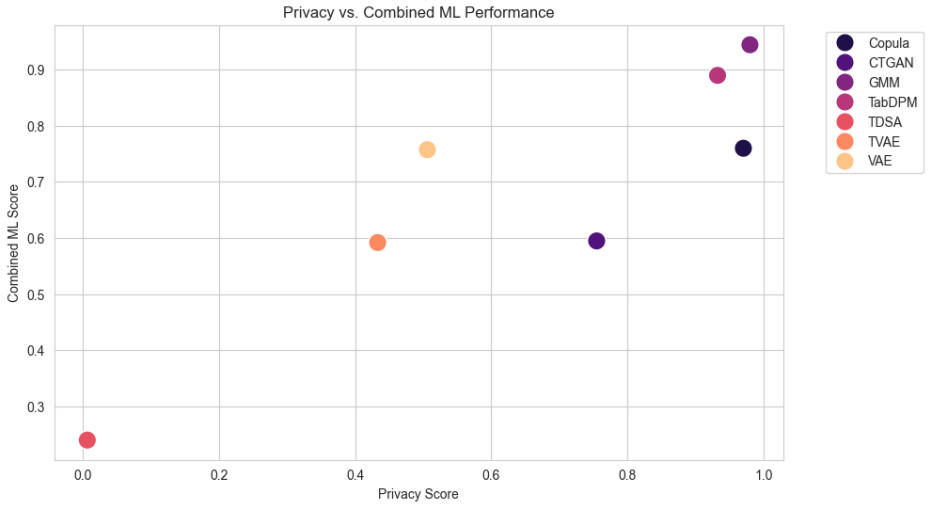

Figure 5 shows the scatter plot of the privacy scores for various synthetic datasets versus their corresponding combined machine learning scores. The combined machine learning score is normalised to fit into a range of 0 and 1 and consist of weighted classification, time-series forecast (use five months of bill and payment amount to predict the sixth month) and clustering scores for five clusters. The weightage for classification to time-series forecast and clustering were 4:3:3, respectively. From Figure 5, it can be observed that only the GMM and TabDPM SDG exceeds the 0.8 score for both combined ML and privacy

At the same time, the classification accuracy of synthetic datasets which is greater than 0.5, 0.6 and 0.8 are 0.79, 0.77 and 0.78 respectively (not shown in Figure 5). While the differences are marginal, it does suggest that there exist inherent trade-offs between privacy and ML performance.

Differential weights for privacy vs ML

Figure 6

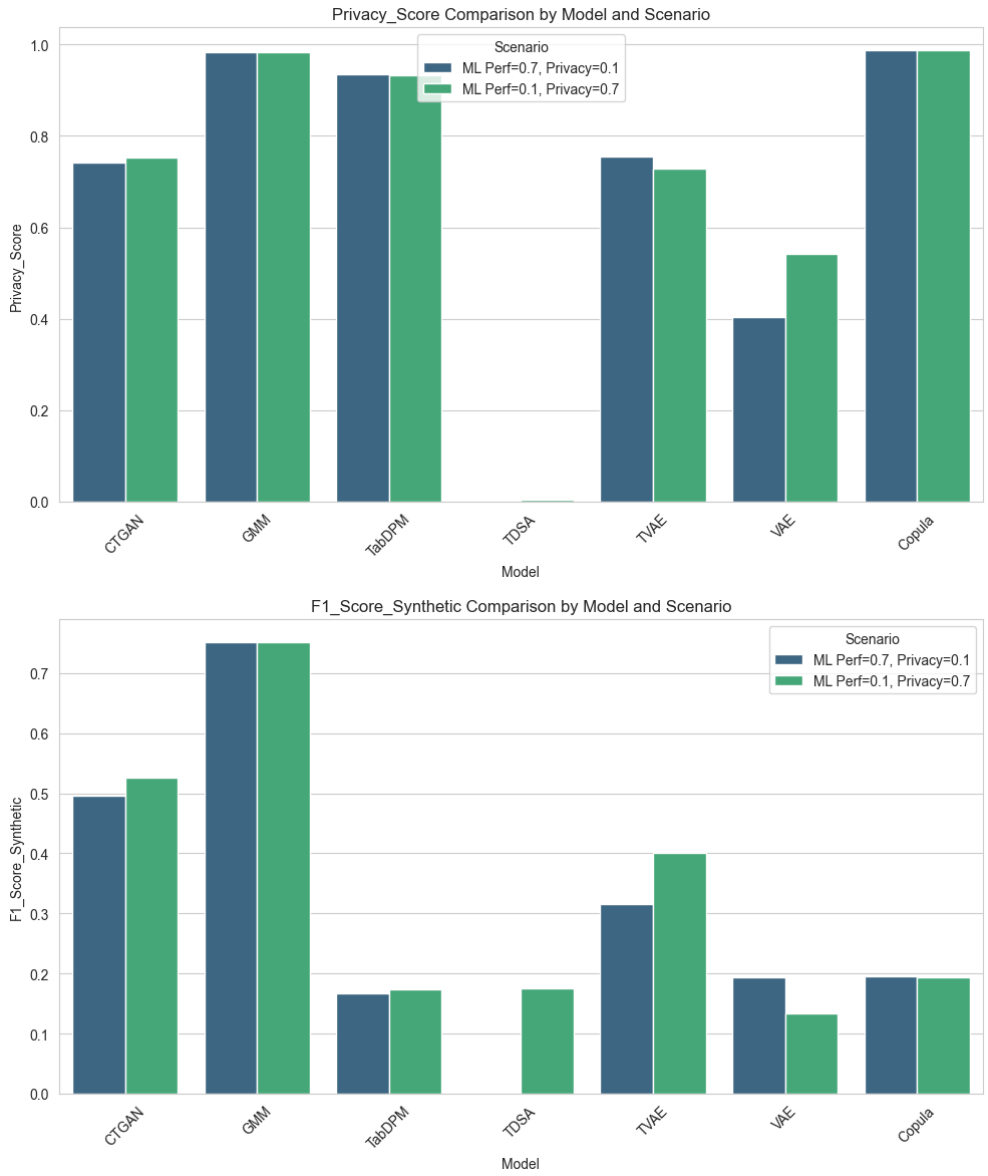

Figure 6 shows the effects of changing the weights for the measures of statistical distance, structural similarity , machine learning and privacy measures. Two scenarios were investigated i.e.

1. Privacy first scenario where weights for measures of statistical distance, structural similarity, machine learning and privacy are 0.1, 0.1, 0.1 and 0.7 respectively
2. Utility first scenario where weights for measures of statistical distance, structural similarity, machine learning and privacy are 0.1, 0.1, 0.7 and 0.1 respectively

Prioritising privacy (ML Perf=0.1, Privacy=0.7) leads to result in similar or slightly higher privacy score values for most models. TVAE shows a slight decrease in privacy score when privacy is weighted higher. Prioritizing privacy (and thus lowering the weight for ML performance) leads to a lower or similar f1-score for the synthetic

data (which might reflect ML utility) for models like VAE and GC. However, for CTGAN and TVAE, the F1 score increases slightly when privacy is prioritized. For TDSA, it increases from zero. For GMM and TabDPM, it remains relatively constant.

Privacy prioritisation also led to a higher composite weighted* score for most models (CTGAN, GMM, TabDPM, TVAE, GC) compared to prioritizing ML performance (ML Perf=0.7, Privacy=0.1). The exception is VAE, where the score slightly decreases, and TDSA, where the score increases significantly but remains negative.

Prioritising machine learning (ML Perf=0.7, Privacy=0.1) led to f1-score for synthetic data being *higher or similar* for several models (GMM, TabDPM, VAE, GC) when ML performance is weighted higher. However, this is not universal; CTGAN, TVAE show a *lower* F1 score in this scenario, and TDSA has an F1 score of zero.

Machine learning prioritisation has led to overall composite weighted score* to be *lower* for most models (CTGAN, GMM, TabDPM, TVAE, GC) compared to when privacy is weighted higher. This suggests that, according to the weighting formula used, the gains in ML performance might not fully compensate for the potential reduction in privacy or other factors in the overall score. VAE is a slight exception, showing a marginally higher weighted score, while TDSA shows a drastically lower (more negative) score. The privacy score is often *slightly lower or similar* compared to when privacy is weighted higher. This is expected, as less emphasis is put on the privacy component. TVAE is an exception, showing a slightly higher privacy score in this scenario.

In summary, weighting ML performance higher often results in slightly lower privacy scores and potentially higher F1 scores (though not always), but it tends to decrease the overall composite weighted score for most models in this dataset and can increase computation time for several methods. This highlights the trade-off being evaluated: prioritizing ML performance directly does not necessarily lead to the best overall outcome as defined by the Weighted score metric in this specific evaluation.

Source*: Author own calculation

4. Discussion

The experimental findings highlight the effectiveness of the proposed weighted evaluation framework in assessing synthetic data quality across privacy, fidelity, and utility dimensions. Simpler generative models like the GMM achieved the highest composite final weighted scores, with GMM, TabDPM, and GC all scoring above 0.80. These models produced synthetic datasets that balance statistical similarity, structural similarity, machine learning utility, and privacy preservation well. In contrast, the TDSA method performed poorly due to high statistical divergence from real data, significantly lowering its overall score.

The framework effectively discerns nuanced differences in generator performance, providing a comprehensive measure of quality. It aligns with recent efforts to develop unified metrics for synthetic data evaluation, offering flexibility to emphasize various aspects via weighting.

Fidelity Dimension: Many models preserved global structure well but differed in distributional details. Four of the seven SDGs produced synthetic data with high correlation similarity to real data, implying intact inter-feature relationships. Distributional similarity proved more challenging, with only CTGAN achieving a Kolmogorov–Smirnov distribution similarity above 0.7.

Privacy Dimension: Most methods produced synthetic data with strong privacy guarantees, with GC, GMM, and TabDPM yielding privacy scores exceeding 0.8, indicating low re-identification risk. TDSA had the weakest privacy performance, suggesting its synthetic data retained more identifiable traces of the original data.

Machine Learning Utility: Results were mixed. In classification tasks, five out of seven synthetic datasets outperformed real data on accuracy, hinting that certain generative models can produce synthetic samples easier for classifiers to learn. However, when evaluating classification by the F1-score, the real dataset outperformed synthetic data in most cases, indicating some synthetic data may not capture minority class characteristics or nuanced decision boundaries as effectively as real data.

The framework's ability to aggregate these facets into a single weighted score provides a convenient summary, yet disaggregated scores remain crucial for diagnosing specific strengths or weaknesses of a synthetic dataset.

A key insight is the inherent trade-off between privacy and utility in synthetic data generation. Only GMM and TabDPM achieved both high privacy and high utility simultaneously, suggesting these techniques were most effective at generating data that is both safe and useful. Other methods fell short in one dimension when excelling in the other. For example, CTGAN and TVAE produced strong utility but lower privacy scores, whereas the GC method excelled in privacy at the expense of slight utility reduction.

The weighted evaluation framework quantitatively captures this balance, giving practitioners a transparent view of how much utility is gained for a given privacy cost, supporting informed decisions on which synthetic data method best fits their risk appetite and analytical needs.

The framework's flexibility to reconfigure the importance of each dimension to suit specific use cases is a significant advantage. Ablative experiments revealed how different prioritizations impact the composite final weighted score and highlighted model-specific behaviours. Weighting privacy higher led to higher overall scores for most models compared to the ML-heavy scenario. This suggests that most generators were better at improving privacy than boosting predictive performance.

In summary, the discussion confirms that the generalized weighted framework captures multiple facets of synthetic data fidelity and utility and privacy in a unified manner, providing a tunable mechanism to explore trade-offs between statistical similarity, analytic utility, and privacy. This holistic perspective is essential for informed decision-making on synthetic data adoption in sensitive domains.

5. Extension

Looking ahead, several promising directions can extend this work to enhance the evaluation framework and facilitate its adoption in practice. These extensions aim to

improve the framework's robustness, broaden its applicability, and align it more closely with real-world needs, particularly in central banking data use cases.

Adaptive or Learned Weighting Strategies: Enhancing the framework to make the weighting adaptive is important. The optimal balance between privacy, fidelity, and utility may not be obvious a priori. Future research could explore algorithms that learn weight coefficients based on specific objectives or constraints, such as multi-objective optimization or reinforcement learning to adjust weights until synthetic data meets target criteria. An adaptive weighting system might consider the context of deployment, automatically assigning higher weight to privacy if sensitive personal identifiers are detected or increasing weight on structural similarity if the downstream task is sensitive to data distribution shifts. Studying how weight tuning affects computational efficiency and model selection can inform practitioners about which techniques are best suited under different priorities. A data-driven or context-aware weighting mechanism reduces reliance on expert judgment and makes the evaluation more objective and personalized.

Enhancing Metrics and Privacy Safeguards: Updating the framework to include improved metrics as privacy and data utility research evolves is essential. Exploring alternative privacy metrics or stronger confidentiality guarantees, such as differential privacy parameters, can provide more formal privacy assurance. Integrating measures of model generalizability or fairness ensures synthetic data supports accurate models without introducing bias. Evaluating whether models trained on synthetic data exhibit similar fairness properties as those trained on real data adds another facet to the quality assessment. Using the weighted score as a feedback signal to train or select better synthetic data generators can push the frontier of synthetic data toward higher quality and reliability.

6. Conclusion

In conclusion, the generalized weighted framework opens numerous avenues for further investigation. By pursuing the extensions above, the framework can be applied as a practical decision-support tool for organizations like central banks that are keen to leverage synthetic data. The ongoing exploration of synthetic data in central banking – whether for data sharing, open research, or enhancing AI models with limited real data – stands to benefit from a robust evaluation standard. Our future work will focus on refining this framework, validating it in collaborative settings, and ensuring it remains aligned with both technological advances in synthetic data generation and the evolving requirements of data governance in the financial industry. Through continuous improvement and adaptation, we anticipate that this framework, with its weighting flexibility and multi-dimensional perspective, can become an integral part of how synthetic data's value and risks are assessed in practice, facilitating greater trust and wider adoption of synthetic data solutions in sensitive domains.

References

Caceres, H. E., & Moews, B. (2024). *Evaluating utility in synthetic banking microdata applications* (arXiv:2410.22519). arXiv. https://doi.org/10.48550/arXiv.2410.22519

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Drechsler, J. (2011). *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation* (Vol. 201). Springer. https://doi.org/10.1007/978-1-4614-0326-5

European Union. (2024). *Article 10: Data and Data Governance | EU Artificial Intelligence Act*. https://artificialintelligenceact.eu/article/10/

Goyal, M., & Mahmoud, Q. H. (2024). A Systematic Review of Synthetic Data Generation Techniques Using Generative AI. *Electronics*, *13*(17), Article 17. https://doi.org/10.3390/electronics13173509

Irving Fischer Committee on Central Bank Statistics. (2023). *Guidance Note 3 – Data-sharing practices*.

Jordon, J., Houssiau, F., Cherubin, G., Cohen, S. N., Szpruch, L., & Bottarelli, M. (2022). *Synthetic Data—What, why and how?* Alan Turing Institution.

Kotelnikov, A., Baranchuk, D., Rubachev, I., & Babenko, A. (2024). *TabDDPM: Modelling Tabular Data with Diffusion Models*. https://doi.org/10.5555/3618408.3619133

Lapteva, E. K. (2023). *Utility and confidentiality assessment of synthetic company data*. IFC Satellite Seminar on "Granular data: new horizons and challenges for central banks."

Ling, K. L. F. (2013). *Application of micro-data for systemic risk assessment and policy formulation*. Workshop on Integrated Managemet of Micro-databases, Porto.

Ong, C. A., Escalanda-Lo, C., Gabriel, M., & Reyes, R. (2024). Research for All: Exploring machine learning applications in generating synthetic datasets. *2024 Philippines Statistical Association Annual Conference*. https://events.psai.ph/2024/ac/

Osorio-Marulanda, P. A., Epelde, G., Hernandez, M., Isasa, I., Reyes, N. M., & Iraola, A. B. (2024). Privacy Mechanisms and Evaluation Metrics for Synthetic Data Generation: A Systematic Review. *IEEE Access*, *12*, 88048–88074. https://doi.org/10.1109/ACCESS.2024.3417608

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

Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling Tabular data using Conditional GAN. *Advances in Neural Information Processing Systems*, *32*. https://proceedings.neurips.cc/paper/2019/hash/254ed7d2de3b23ab10936522dd547b78-Abstract.html

Yeh, I.-C., & Lien, C. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, *36*(2, Part 1), 2473–2480. https://doi.org/10.1016/j.eswa.2007.12.020

Generalised Weighted Framework for Synthetic Data Evaluation[1]

Chiung Ching Ho BNM

1 This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, Bank Negara Malaysia, the BIS, the IFC or the other central banks and institutions represented at the event.

Overview

Privacy, utility and fidelity for trust-worthy synthetic data

There is a need to use synthetic data by Central Banks
driven by both legislation and a need to maintain privacy of data

Synthetic data produced by Central Banks needs to be trusted
to fulfil privacy, utility and fidelity (PUF) requirements

An assessment framework is proposed to evaluate synthetic data
produced by synthetic data generators (SDG) that assesses PUF requirements using a flexible and extensible framework

Index

Generalised Weighted Framework for Synthetic Data Evaluation

Motivation : Inspired by an attempt to create a universal metric for evaluation of synthetic data [1], a new framework is proposed that will allow for flexible assessment of synthetic data from a privacy, utility and fidelity (P.U.F) perspective

Problem statement : Data created by SDG needs to be evaluated for privacy, utility and fidelity (PUF) to ensure that synthetic data is credible to be shared or published by central banks or authorities, and current solutions does not allow for flexibility in measuring SDG that balances P.U.F needs

Research Question (RQ) 1 : Which of current SDG techniques exhibit good PUF scores?

Research Question (RQ) 2 : How do we balance P.U.F measures for different SDG to address different analytical needs?

Assessing synthetic data

- Synthetic data is data that has been generated using a purpose-built mathematical model or algorithm, with the aim of solving a (set of) data-science task(s)[2]
- Synthetic data is being used by central banks to enable data sharing without compromising on privacy or infringing on legislation[3]
- Reasons for sharing data using synthetic data in central banks includes
  - Sharing micro data for research purposes[4]
  - Justification of data collection rigor [5]
  - Data disclosure risk mitigation and[6]
  - Sharing of granular data instead of aggregated data[7]
- Examples of synthetic data shared by central banks includes micro data from surveys, non-financial firm-level data, and loans to legal persons[4][5][6][7]

Synthetic data generation (SDG) techniques comparison

Name of Technique	Strengths	Weaknesses
Gaussian Diffusion Models (GDM)	High flexibility in modelling distributions Good for high-dimensional data	Computationally intensive Requires large amounts of training data
Gaussian Copula (GC)	Maintains statistical properties Suitable for tabular data	Assumes a Gaussian dependence structure Struggles with complex, non-linear relationships
Conditional Tabular GAN (CTGAN)	Handles imbalanced and multi-modal data Captures complex dependencies	Training instability Sensitive to hyperparameter tuning
Tabular Variational Autoencoder (TVAE)	Effective in capturing latent structures Suitable for tabular data	May overfit on small datasets Needs careful parameter tuning
Gaussian Mixtures Models (GMM)	Models data as a combination of distributions Interpretable and straightforward	Assumes data can be represented as Gaussian mixtures Sensitive to initialization
Time Dependent Self Attention (TDSA)	Strong at time-series data generation Preserves temporal correlations	Limited to time-series data Computationally expensive for large datasets
Tabular Diffusion Probabilistic Model (TabDPM)	Robust for tabular data with complex dependencies Flexible with varying distributions	Computationally heavy Requires careful parameter configuration

Table 1 : A comparison of SDG techniques

Assessing synthetic data : P.U.F for credible synthetic data

Privacy

- ❑ Is x in the dataset?
- ❑ Can unknown data about x be found?
- ❑ Can I recreate the original data?

Utility

- ❑ How does the synthetic data perform as compared to the original data in specific task?

Fidelity

- ❑ How truthful is the synthetic data?
- ❑ How statistically similar is the synthetic data?
- ❑ How similar is the distribution of the synthetic data?

There is always a trade-off between privacy and utility, while fidelity may proxy for utility

Generalised Weighted Framework for Synthetic Data Evaluation

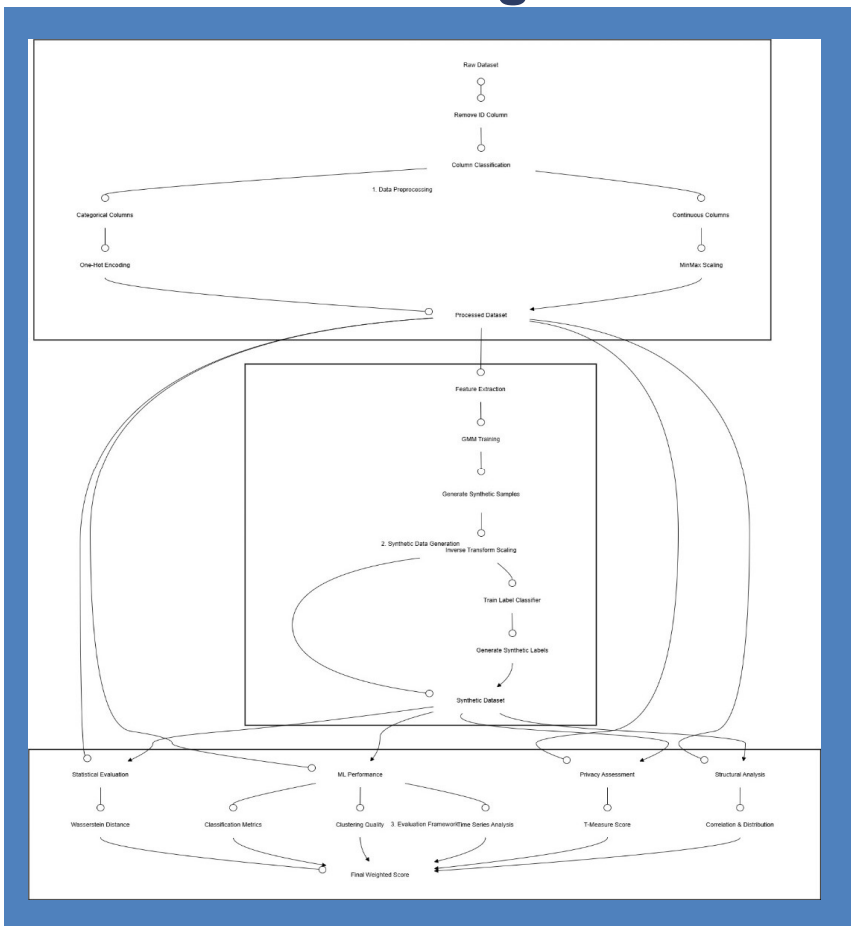

Figure 1: Synthetic data evaluation process flow

Phase 1 - Data Preprocessing
- **Start with the raw dataset**
- **Remove any ID columns**
- **Classify columns into two types:**
  - Categorical (like SEX, EDUCATION, MARRIAGE, PAY status)
  - Continuous (like AGE, BILL amounts, PAYMENT amounts)
- **Process each type differently:**
  - Categorical columns get one-hot encoded
  - Continuous columns get scaled using MinMax scaling
- **Result: A cleaned, processed dataset**

1.Preprocessing is dependent on data set

Phase 2 - Synthetic Data Generation
- **Take the processed dataset**
- **Extract features for GMM training**
- **Train the Gaussian Mixture Model**
- **Generate synthetic samples**
- **Convert the synthetic data back to original scale**
- **Generate appropriate labels using a classifier**
- **Result: A synthetic dataset that mimics the original**

2. The synthetic data generator (SDG) is chosen based on use case

Phase 3 - Evaluation
- **Compare real and synthetic data across four aspects:**
  - Statistical similarity (using Wasserstein distance)
  - Machine learning performance (classification accuracy, clustering quality, time series analysis)
  - Privacy assessment (using T-measure score)
  - Structural analysis (correlations and distributions)
- **Combine all metrics into a final weighted score**

Each metric contributes equally (25%) to the final evaluation score, which helps assess how well the synthetic data captures the important characteristics of the original dataset while maintaining privacy.

3. Privacy, Utility (ML scores) and Fidelity (statistical and structural) measures can be changed or weighted as needed

Figure 2: Synthetic data evaluation measures using Gaussian Mixture Models as a Synthetic Data Generator

Experiment using credit card data

The synthetic data was created from an open source credit card dataset*
This dataset was chosen as it was suitable for assessing utility through machine learning applications
It has 30,000 entries (excluding header) containing information about credit card clients. Here are the key features:

Financial

- LIMIT_BAL: Credit limit balance
- BILL_AMT1 through BILL_AMT6: Bill amounts for six consecutive months
- PAY_AMT1 through PAY_AMT6: Payment amounts for six consecutive months

Demographics

- SEX: Gender of the client
- AGE: Age of the client
- EDUCATION: Education level
- MARRIAGE: Marital status

Historical payment

- PAY_0 through PAY_6: Payment status history for seven months
- "Default payment next month": Binary indicator (0 or 1) predicting whether the client will default on their payment in the next month

* Yeh, I.-C., & Lien, C. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, *36*(2, Part 1), 2473–2480. https://doi.org/10.1016/j.eswa.2007.12.020

Fidelity : Measuring Wasserstein distance and data correlation & distribution

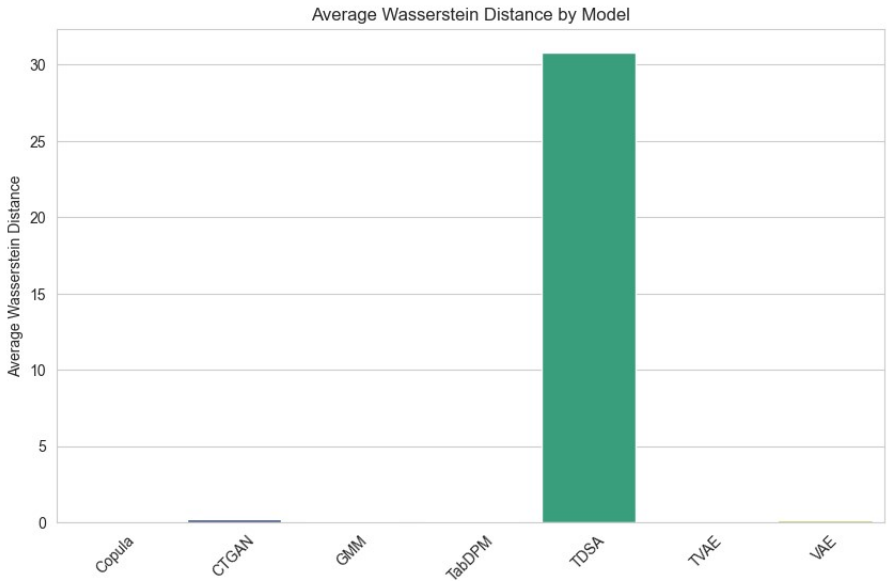

Figure 3 : Average Wasserstein distance for all columns for synthetic data relative to original data except for TDSA

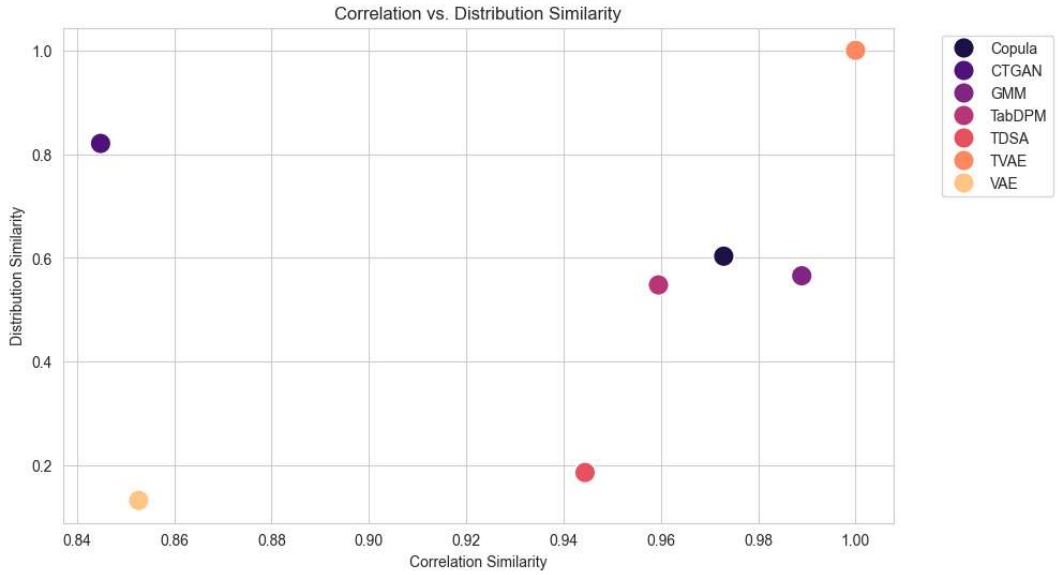

Figure 4 : TVAE generated synthetic data has high correlation and distribution similarity

Utility : Classification and Clustering and Time Series Forecast

Figure 5 : Overall weighted scores for synthetic data produced by SDGs are similar except for TDSA

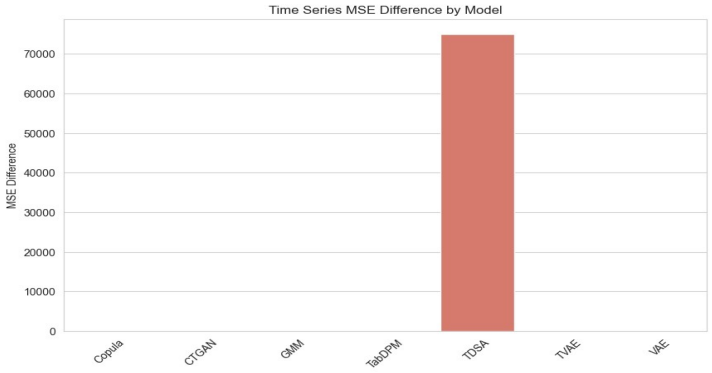

Figure 6: Time series forecast mean square error for synthetic data produced by various SDG with TSDA as an outlier

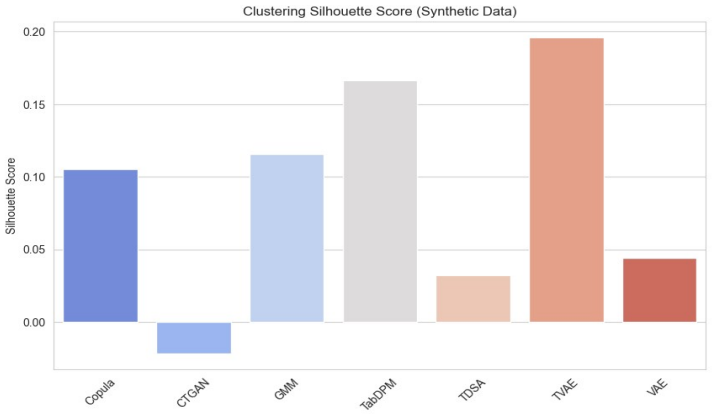

Figure 7: Clustering scores for synthetic data produced by various SDGs are generally poor

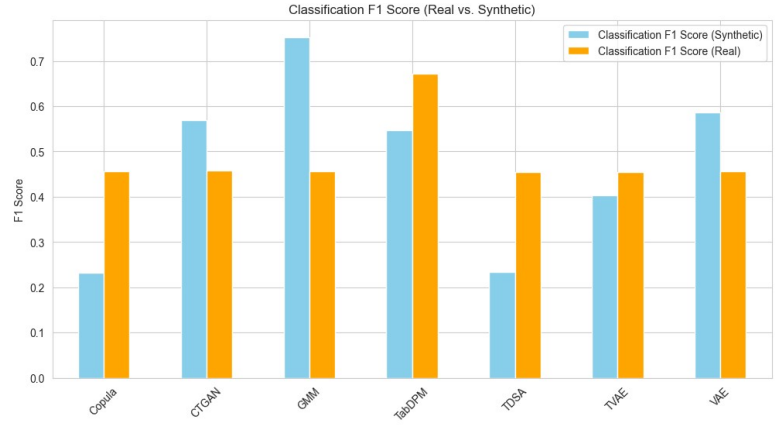

Figure 8: F1 scores for synthetic data produced by SDGs

Privacy : T-measure scores for various synthetic data produced by various SDG

T-measure is the absolute difference between the synthetic and original column mean for all numeric columns, normalised by the standard deviation. This is then transformed into the privacy score

Figure 9: Privacy scores for datasets produced by various SDG with TDSA as an outlier

Privacy vs Utility trade-off : Choosing the SDG that fulfils both

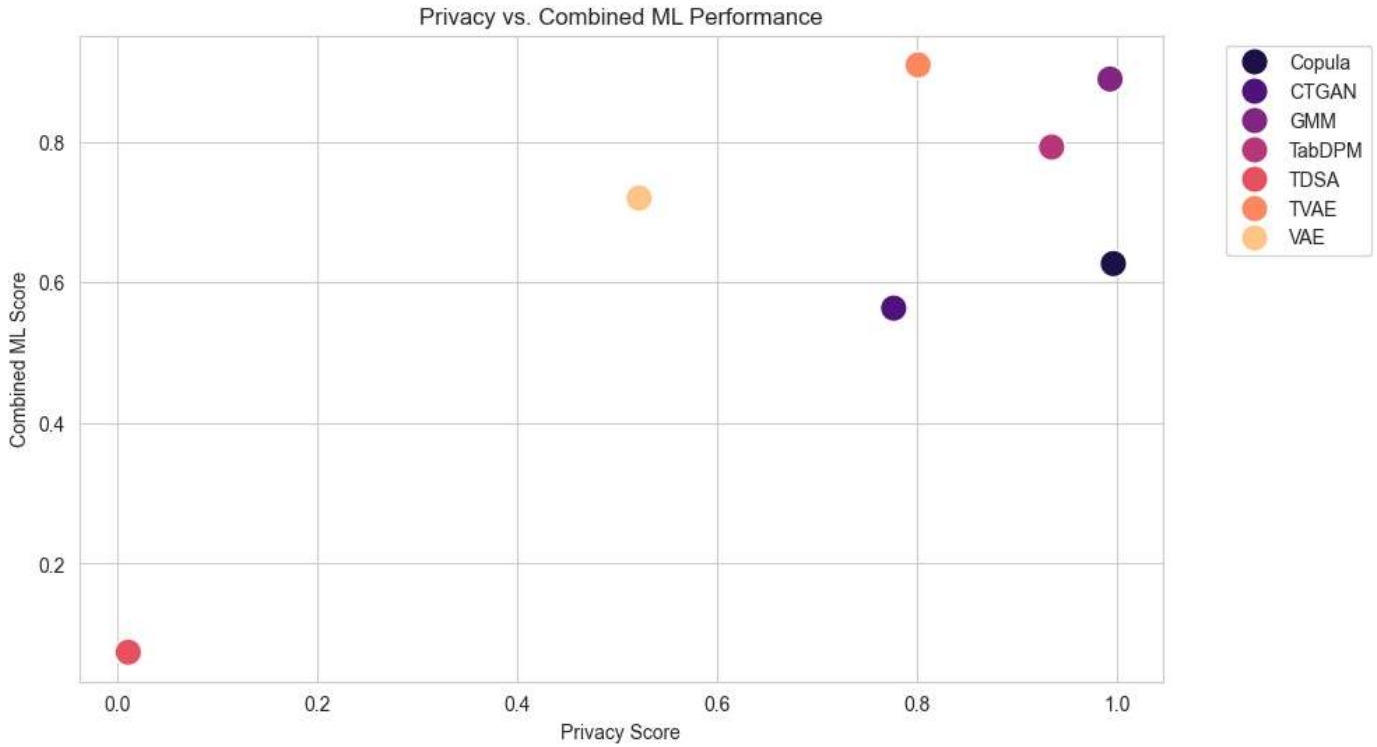

Figure 10: Privacy vs Utility (Combined ML) Score plot

Discussion : Weighted score aligns with privacy, utility and fidelity measures

Model	Weighted Score	Privacy	Utility	Fidelity	Runtime(s)	Evaluation Summary
GMM	0.89	High	High	Moderate	34.5	Best overall performer. High privacy and utility and moderate fidelity. Highest weighted score.
TVAE	0.86	High	High	High	1113.2	High privacy, high utility, and high fidelity. Weighted score reflects this balance.
TabDPM	0.83	High	Moderate	Moderate	365	High privacy with moderate utility and fidelity. Weighted score reflects this balance.
Copula	0.82	High	Moderate	Moderate	84448.3	High privacy, but moderate utility and fidelity. Weighted score reflects this balance.
CTGAN	0.77	Moderate	Low	High	139.3	Moderate privacy, low utility but high fidelity. Weighted score aligns with this.
VAE	0.65	Low	Moderate	Low	97.5	Low privacy, moderate utility and low fidelity. Weighted score aligns with this.
TDSA	-7.19	Low	Low	Low	2367.6	Poor in all dimensions. Negative weighted score reflects its poor performance.

Table 2 : Weighted scores and PUF scores alignment

Conclusion

Summary

- The proposed framework is generalisable to support different measures for different analytics tasks, answering RQ 1 & 2.
- Preprocessing of data is however difficult to generalise as it is very task-dependent.
- Future expansion of the framework may include support for non-structured data, federated machine learning and addition of differential privacy mechanism

Recommendations

1.Privacy-Critical Applications:

- ❑ **GMM**: Best for high privacy, fidelity, and minority class preservation. Validate to avoid overfitting
- ❑ **Copula**: Use for analysis in sensitive domains; at a cost of long processing times

2.Utility-Focused Applications:

- ❑ **GMM**: Best for high privacy, fidelity, and minority class preservation. Fast processing times
- ❑ **TVAE**: Strong for utility and privacy in sensitive domains

3.Avoid:

- **TDSA**: Poor in all metrics; not recommended for imbalanced datasets and datasets with non-sequential data

References

[1] Chundawat, V. S., Tarun, A. K., Mandal, M., Lahoti, M., & Narang, P. (2024). TabSynDex: A Universal Metric for Robust Evaluation of Synthetic Tabular Data (arXiv:2207.05295). arXiv. https://doi.org/10.48550/arXiv.2207.05295

[2] European Union. (2024). Article 10: Data and Data Governance | EU Artificial Intelligence Act. https://artificialintelligenceact.eu/article/10/

[3]Jordon, J., Houssiau, F., Cherubin, G., Cohen, S. N., Szpruch, L., & Bottarelli, M. (2022). *Synthetic Data—What, why and how?* Alan Turing Institution.

[4] Bender, S., & Staab, P. (2015). *The Bundesbank's Research Data and Service Centre (RDSC)—Gateway to treasures of micro data on the German Financial System*. IFC workshop on "Combining micro and macro statistical data for financial stability analysis. Experiences, opportunities and challenges".

[5] Caceres, H. E., & Moews, B. (2024). *Evaluating utility in synthetic banking microdata applications* (arXiv:2410.22519). arXiv. https://doi.org/10.48550/arXiv.2410.22519

[6] Lapteva, E. K. (2023). *Utility and confidentiality assessment of synthetic company data*. IFC Satellite Seminar on "Granular data: new horizons and challenges for central banks".

[7] Ong, C. A., Escalanda-Lo, C., Gabriel, M., & Reyes, R. (2024). Research for All: Exploring machine learning applications in generating synthetic datasets. *2024 Philippines Statistical Association Annual Conference*. https://events.psai.ph/2024/ac/

Addendum : Reflection on generative AI (GenAI) tools for enabling research (as of 31.12.2024)

	Research	Coding	Writing	Analysis
Helps	Generate new ideas	Speed up coding	Conciseness and tone	Quick initial comparisons
Caveats	Hallucination and recency	LLMs are throttled LLMs needs management	Loss of personality	Surface level analysis
Future hopes	Integrated into citation management	More efficient models to enable self-service coding	Increased support for academic writing	Increased evidence of reasoning