Irving Fisher Committee on
Central Bank Statistics

BIS

IFC-Bank of Italy Workshop on "Data science in central banking: enhancing the access to and sharing of data"

17-19 October 2023

Research for all: exploring machine learning applications in generating synthetic datasets[1]

Carmelita Esclanda-Lo, Gabriel Masangkay,
Chelsea Anne Ong and Rossvern Reyes,
Bangko Sentral ng Pilipinas

[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, the BIS, the IFC or the other central banks and institutions represented at the event.

Research for All: Exploring machine learning applications in generating synthetic datasets

Carmelita Esclanda-Lo, Gabriel Masangkay, Chelsea Anne Ong, Rossvern Reyes[1]

Abstract

Sharing granular data with internal and external parties is becoming more challenging for central banks due to stricter laws, regulations, and growing concerns about data privacy. These factors hinder or slow the progress of collaborative and data-centric research and innovation. However, with the advancements in data generation, it is now possible to generate synthetic datasets, which mimic the statistical characteristics of the actual data but mask or hide any private information contained therein, using artificial intelligence (AI) methods.

In this paper, we compare the results of several AI methods, *i.e., the Synthetic Minority Over-sampling Technique (SMOTE), Generative Adversarial Networks (GANs), and Tabular Variational Autoencoders (TVAE)*, in generating synthetic granular and tabular datasets, using the 2023 Quarter 1 Consumer Expectation Survey (CES) dataset of the Bangko Sentral ng Pilipinas (BSP).

The results demonstrate that TVAE generates synthetic tabular data with the highest fidelity, privacy, and utility among the other algorithms tested. Conversely, GANs performed poorly in terms of data fidelity and utility.

Keywords: Synthetic Data, Artificial Intelligence, Machine Learning, Data Sharing

JEL classification: C45 C81

[1] This paper was prepared for the IFC and Bank of Italy Workshop on "Data Science in Central Banking" which was held on 17-19 October 2023 in Rome, Italy. Carmelita Esclanda-Lo (esclandacg@bsp.gov.ph), Gabriel Masangkay (masangkayga@bsp.gov.ph), Chelsea Anne Ong (ongcs@bsp.gov.ph), and Rossvern Reyes (reyesros@bsp.gov.ph) are from the Department of Economic Statistics of the BSP. The views expressed are those of the authors' only and not those of the BSP.

Table of Contents

1. Introduction

Background

In advancing data-centric research and fostering innovation, central banks are progressively embracing the idea of sharing granular data with researchers by launching data-sharing initiatives, which allow data sharing under conditions that protect sensitive information with confidentiality. Currently, data-sharing frameworks are established to protect sensitive information. Still, these processes are often tedious and stringent, burdening data producers (central banks) and requesters due to strict data privacy and confidentiality regulations.

To address data-sharing concerns, the BSP developed its Data Governance Manual, which specifies, among many provisions, the rules on sharing data with external parties and protecting the privacy of its data subjects. As part of its data governance initiatives, the BSP crafted its data masking guidelines outlining rules on obfuscating personal identifiable information (PII) and sharing data with external parties.

Other Philippine government agencies also have methods of disseminating data to the public. Most notably, the Philippine Statistics Authority (PSA) makes the microdata from their surveys available through the PSA Data Archive. As part of the request, data requesters are asked to submit their application to the agency head stating the purpose of their request. The PSA will provide a subset of the requested microdata upon its approval. For some data, access could only be done via onsite access to the PSA’s Data Enclave Center.

The emergence of synthetic data through AI offers a potential solution for improving data sharing and addressing these data-sharing concerns. Synthetic datasets produced by AI exhibit identical mathematical or statistical characteristics as their original counterparts while preserving privacy.

Objectives

This study aims to explore and compare various machine learning (ML) methodologies to generate another dataset with mathematical and statistical structures similar to the original dataset. More precisely, this research aims to evaluate the performance, check the limitations, and dissect the practical applicability of various AI-based models, including Gaussian Mixture Models (GMM) and advanced Generative AI models, such as GANs and Variational Autoencoders (VAE), in creating synthetic granular and tabular data that ensures utility and privacy. In addition, this study aims to show a preliminary pipeline for implementing and evaluating the said ML models.

Importance of Synthetic Data Generation in Research and Applications

Assefa et al. (2020) highlighted the importance of effective synthetic data generation in the financial sector. Since financial data includes PII, data sharing is highly restricted. Synthetic data generation addresses the challenges of publishing stream data by enhancing data privacy and security, facilitating model development and testing, and overcoming actual financial data unavailability. Furthermore, it discusses the complexities of producing realistic synthetic financial data, including the need for data generators to accurately capture the underlying patterns, correlations, and distributions of actual financial data. Addressing these challenges is essential to ensure the utility of synthetic data. For central banks, adopting synthetic data can be a strategic solution for promoting research collaborations while preserving data privacy.

2. Synthetic Data Generation Methodologies

Methodologies in Tabular Synthetic Data Generation

Pathare et al. (2023) generated synthetic data on unbalanced, balanced, numerical-only datasets, categorical-only datasets, and datasets with a mix of numerical and categorical attributes using multiple models, i.e., Conditional Tabular Generative Adversarial Network and Classification and Regression Trees (CART), with results based on different comparison parameters which are accuracy, propensity score, log-cluster, and execution time. For all types of datasets analyzed in the study, CART generates data with the highest quality, while Bayesian networks performed the worst among other models.

One of the overarching objectives of generating synthetic datasets is to protect the PII of data subjects while maintaining the usefulness or utility of the data. Little et al. (2021) assessed the disclosure risk of the synthetic data they generated through CART and GANs. They found that the synthetic data produced via CART had the highest utility but also had the highest risk of disclosure. On the other hand, Table GAN synthetic data produced the lowest risk but also had the lowest utility. This finding corroborates an important concept regarding utility and risk as mentioned in the United Nations Economic Commission for Europe's (UNECE) (2022) Synthetic Data for Official Statistics: *a trade-off exists between utility and disclosure risk*, i.e., as the utility of the synthetic data increases, the disclosure risk increases exponentially.

Use of Synthetic Data Generation in Central Banks

The integration of synthetic data methodologies in central banking is in its nascent stage. Synthetic data generation was mentioned in the first IFC report on data sharing (IFC, 2015), where it was cited that synthetic data could *"transform the original confidential microdata into artificial microdata with the same statistical properties that third parties can use."*

On the other hand, National Statistics Offices (NSOs) worldwide have developed and operationalized synthetic datasets for various use cases (UNECE, 2022). Some NSOs have developed synthetic data to improve the efficiency of data-sharing processes. For instance, Statistics Canada developed a version of a census-based database for testing and running the new dynamic micro-simulation model of the Canadian retirement and income system. The synthetic database would enable policymakers to experiment and model changes to the Canada Pension Plan. The micro-simulation model would use the original data for the final analysis.

3. Data

We used the First Quarter 2023 Consumer Expectation Survey (CES) dataset, formatted in Microsoft Excel, to generate synthetic datasets. The CES, which the BSP conducts, is a nationwide quarterly survey on consumers' sentiments, i.e., Philippine household sentiments, on family income, financial situation, and economic condition of the country for the current quarter, next quarter, and next 12 months. With over 1,000 columns or variables, which reflects the diverse sets of survey questions, pre-processing was needed to transform the CES dataset into a dataset that can be used with AI methods.

Sample CES variables

A total of 36 variables is analyzed in this study

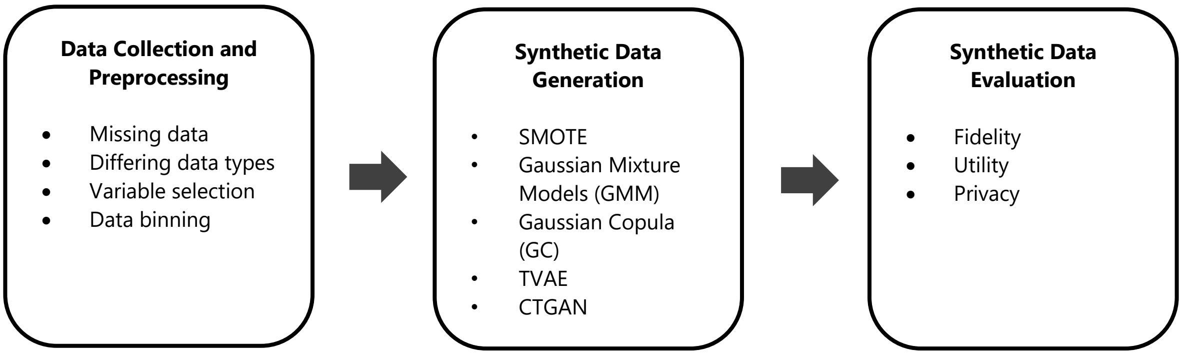
Table 1

Variables	Description	Values
AGE	Age	0 - 100
INCOME	Income Group	Low, Middle, High
SEX	Sex	Male, Female
C5C	Inflation Rate in the Current Quarter	Less than 0%, 0.1% - 1.9%
E1S	Has Family Savings	Yes, No
B1S	Present Family Situation	Better, Same, Worse

4. Methodology

Methodology Overview for Synthetic Data Generation

Pipeline for generating and evaluating synthetic datasets. Figure 1



To achieve our objectives, we created a pipeline for generating and evaluating synthetic datasets, as shown in Figure 1. We first collected and pre-processed the CES dataset, and selected 36 variables based on a simulated research problem on identifying the determinants of households' inflation outlook. Subsequently, we generated the synthetic datasets through various statistical and generative machine learning models using open-source Python Data Synthesizer libraries (e.g., Synthetic Data Vault (SDV) and YData Synthetic). Lastly, we evaluated the generated synthetic datasets using metrics based on three key dimensions or qualities: fidelity, utility, and privacy.

Data Collection and Processing

With reference to the study of Basilio (2010) on the determinants of households' inflation outlook, we limited the number of variables of the CES from over 1,000 to 36 variables for computational efficiency. As listed in Appendix A, these variables include the respondents' demographic characteristics and outlook on various economic and financial indicators.

After reducing the variables, we cleaned and pre-processed the data. We then transformed columns containing more than one data type (e.g., a combination of string, float, integer) to its proper data type. Lastly, we binned the age variable to reduce the effects of outliers.

Synthetic Data Generation Libraries

We used open-source Python libraries, such as the SDV and YData Synthetic libraries, to generate synthetic datasets.

The Synthetic Data Vault (SDV)

The SDV is an ecosystem designed by DataCebo, Inc. for synthetic data generation and evaluation using ML (i.e., ranging from classical statistical to deep learning methods). By using SDV, we could define the constraints for pre-processing the selected variables and compare the resulting synthetic datasets with the original.

In the SDV implementation, metadata creation is required where the data type of each variable must be explicitly specified. Then, the generated model creates samples of synthetic data that retain the format and the mathematical properties of the original or actual dataset. In addition, SDV has a conditional sampling feature that allows the generation of hypothetical scenarios by fixing values to extreme cases or imputing data.

YData Synthetic

YData Synthetic is an open-source Python package that generates synthetic tabular and time-series data using classical and state-of-the-art generative models.

Like in SDV, we defined first the numerical and categorical columns before fitting the data to the model. A sample of the synthetic data is then generated based on the size of the actual dataset. The models chosen were CTGANs and GMMs.

Algorithms for Synthetic Data Generation

We utilized and compared the performance of five algorithms in generating synthetic data of the 36 variables from the CES dataset. These algorithms are discussed below.

Synthetic Minority Over-Sampling Technique

SMOTE is an oversampling approach where new instances of the minority class are created by joining all or any of the *k* minority class nearest neighbors to balance the dataset (Chawla, 2002). Suppose we need to double the size of the dataset. In that case, only two neighbors out of the k-nearest neighbors are chosen, and a synthetic sample is created at a random point somewhere between two examples in the feature space.

Gaussian Copula (GC)

GC relies on the fact that separate distributions can be modeled from the joint distribution. Mathematically, this works by applying the Probability Integral Transform, a mathematical method used to transform any random variable with a cumulative distribution function (CDF) into a uniform distribution.

Although the GC algorithm is designed for only numerical data, this synthesizer converts other data types using Reversible Data Transforms (RDTs). Some controls in synthetic data can be made by adjusting the parameters, such as setting the minimum or maximum boundaries, rounding off the values, and setting the distribution shape of numerical columns similar to the actual data.

Conditional Tabular Generative Adversarial Networks (CTGAN)

CTGAN is a variation of the GAN-based method for modeling tabular data. GAN deep learning models have two main components (a Generator and a Discriminator). They are trained through an adversarial learning process where the generator and discriminator compete and improve iteratively. New data is generated by inputting random noise into the generator, introducing interesting variability to the synthetic data.

This algorithm works ideally for data with complete values. For the SDV implementation, in addition to the parameters of GC models, the CTGAN synthesizer has other parameters like the number of epochs or the number of times to train the GAN to improve the model, batch size, learning rate, and embedding dimensions.

Tabular Variational Autoencoder (TVAE)

A tabular variational autoencoder (TVAE) is a mathematical model for compressing and representing tabular data. It consists of an encoder that transforms the input data into a distribution characterized by mean (μ) and standard deviation (σ) parameters, introducing a degree of randomness. During training, the model minimizes an objective function that encourages accurate reconstructions and diverse representations. This approach enables the TVAE to learn a flexible and probabilistic encoding of tabular data, allowing for nuanced understanding and generation of meaningful representations while considering uncertainties in the information.

Like CTGANs, the TVAE works best for data with complete values. Parameters specific to TVAE are the batch size, hidden layer size, regularization, and loss factor.

Gaussian Mixture Models (GMM)

GMMs assume that the data is represented as a mixture of Gaussian distributions characterized by its mean and covariance. GMMs generate synthetic data via sampling from the learned distribution, ensuring that the new data has the same characteristics as the actual data.

Evaluation Metrics

Data Fidelity

Data fidelity pertains to how well synthetic data captures the information in the real dataset. We followed the framework of Platzer and Reutterer (2021) to measure data fidelity, wherein distributions and correlations are measured and compared for synthetic and actual or real data.

- Histogram Plots and Bray-Curtis Similarity Score

Histogram plots were generated for real and synthetic datasets to illustrate the visual differences. An excellent example of a synthetic dataset is one that keeps the shape and distribution close to the original dataset.

The Bray-Curtis Distance is computed to quantify the visual differences between the histogram plots. This measure is used in biology to measure the distances of two compositions between two sites. In this case, these sites are the real and the synthetic data. The Bray–Curtis similarity score is bounded between 0 and 1 and is obtained by the following equation:

$$Bray - Curtis\ similarity = \frac{\sum_{i=1}^{n} \parallel A_i - B_i \parallel}{\sum_{i=1}^{n} \parallel A_i + B_i \parallel}$$

where A_i and B_i represent vectors A and B, respectively. A higher Bray-Curtis similarity score indicates that the count distribution of the synthetic data is closer to the actual data.

- Correlation and Cosine Similarity Score

Meanwhile, the correlation coefficients between the variables were computed to determine whether relationships within the original dataset were preserved. To visually depict the correlation of variables, we plotted the Cramér's V values on the heatmap.

Cramér's V is an extension of Pearson's chi-squared test for independence. It assesses the strength of association or correlation between two categorical variables in a contingency table. A zero value indicates no association, and a one indicates perfect association between the variables. It is calculated by taking the square root of the chi-squared statistic, which is then divided by the product of the sample size and one less than the minimum dimension:

$$V = \sqrt{\frac{\delta^2}{\min{(k-1, r-1)}}} = \sqrt{\frac{\chi^2/n}{\min{(k-1, r-1)}}}$$

where δ^2 is the phi coefficient, $\chi^2$is the chi-squared statistic, *n* is the number of observations, *k* is the number of columns, and *r* is the number of rows.

Additionally, the cosine similarity scores were computed to assess the similarity between the Cramer's V values of the real and synthetic data by computing the cosine of the angle between the two vectors, which fall from -1 and 1. A value closer to 1 denotes higher similarity. The following equation defines cosine similarity:

$$Cosine\ similarity\ (\cos\theta) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} = \frac{\mathrm{A} \cdot \mathrm{B}}{\parallel \mathrm{A} \parallel \parallel \mathrm{B} \parallel}$$

where $\mathrm{A} \cdot \mathrm{B}$ represents the dot product of vectors A and B, while $\parallel \mathrm{A} \parallel$, $\parallel \mathrm{B} \parallel$ represent the Euclidean norm (magnitude) of vectors A and B, respectively. A higher cosine similarity score indicates that the distribution and correlation of the synthetic data are closer to the actual data.

- Statistical Similarity Score

To account for the combined histogram and correlation scores, we computed the statistical similarity score as the average of the histogram Bray-Curtis similarity score and the cosine similarity score of the Cramer's V values.

Data Utility

Synthetic data generation methodologies such as removing information and adding noise can reduce the usability of the data. However, the generated synthetic data should yield an ML model performance similar to the actual data. To this end, we simulated the research problem of Basilio (2010), which identified the determinants of households' inflation outlook.

We developed a multi-class classifier to predict the range of inflation rates in the next 12 months given in the CES data. Python's PyCaret package, an open-source library that automates ML workflows, was utilized for simplicity and ease of implementation.

The algorithms used were K-nearest neighbors, tree-based models (e.g., decision tree, random forest, gradient boosting method), linear models (e.g., logistic regression, linear support vector machines, ridge), and Bayesian models (e.g., Naïve Bayes, Bernoulli). We further improved the top-performing ML model by hyperparameter tuning and assessed this through the Area Under the Receiver Operating Characteristic Curve (ROC AUC) score, more commonly called the Area Under the Curve (AUC). Furthermore, we supplemented the results by getting the model's feature importance score.

Data Privacy

Carlini et al. (2021) discussed membership inference attacks in the context of privacy and ML, aiming to determine whether specific actual data points are part of a model's training dataset, which may potentially compromise privacy. One key metric discussed in the paper is the balanced attack accuracy, which assesses how often an attack correctly predicts membership on a combined dataset where each data point is labeled as either real (0) or synthetic (1). Despite its widespread use in various papers, the accuracy metric has limitations. Since it is an average-case metric, it fails to consider the costs of incorrect predictions. For institutions like central banks, having incorrect predictions could expose the actual data, which poses security and privacy risks. Another evaluation metric is the AUC score. An AUC score near 50% indicates random guessing, suggesting that the model cannot effectively distinguish between the actual and synthetic data, thereby preserving privacy. However, this approach averages all false-positive rates, even in high error rates. As a result, the likelihood of revealing the real data is higher. To assess data privacy more effectively, the study discussed using precision, which represents the percentage of true positives among all predicted positives. A high precision score reduces false positives while simultaneously increasing true positives.

Hence, we thoroughly evaluated our synthetic data using the three metrics: accuracy, AUC score, and precision. Accuracy assesses the overall correctness of the model, AUC measures the model's ability to distinguish between classes, and precision focuses on minimizing false positives. Collectively, these metrics aim to ensure data privacy, especially in scenarios where accurately identifying the actual data points from synthetic ones is essential.

Best Synthetic Data Evaluation Criteria

Synthetic Data Evaluation Matrix

Scores: 3 – Excellent, 2 – Fair, 1 - Poor

Table 2

	Data Fidelity	Data Utility	Data Privacy	Decision
3 – Excellent	Has a Statistical Similarity Score of 0.95 or higher	Overall average difference of 0.05 or less	Privacy Score is higher than 0.9	Synthetic data can be used for research.
2 – Fair	Statistical Similarity Score higher than 0.90 but lower than 0.95	Overall difference of 0.05 to 0.10	Privacy Score is higher than 0.8 but lower than 0.9	Synthetic data can be used for research but with conditions
1 – Poor	Has a Statistical Similarity Score of 0.90 or lower	A difference of more than 0.10 Other conditions not satisfying any of the above	Privacy Score is lower than 0.8	Not for research use. Test another algorithm

[1] This decision matrix is devised for evaluating the best synthetic data.

Sources: asq.org/quality-resources/decision-matrix

Table 2 shows a decision matrix for ranking the algorithms regarding data fidelity, utility, and privacy to determine the best algorithm for generating synthetic data. The rubric ranges from one to three, with three indicating excellent performance in the corresponding evaluation criteria.

5. Results

This section evaluates the synthetic data generated from the CES dataset using various open-source Python Data Synthesizer libraries (e.g., SDV, YData Synthetic, and SMOTE). Using machine learning, we demonstrated the utility of synthetic data for research and provided a quantitative evaluation focusing on the fidelity, utility, and privacy of the generated synthetic data.

Data Fidelity

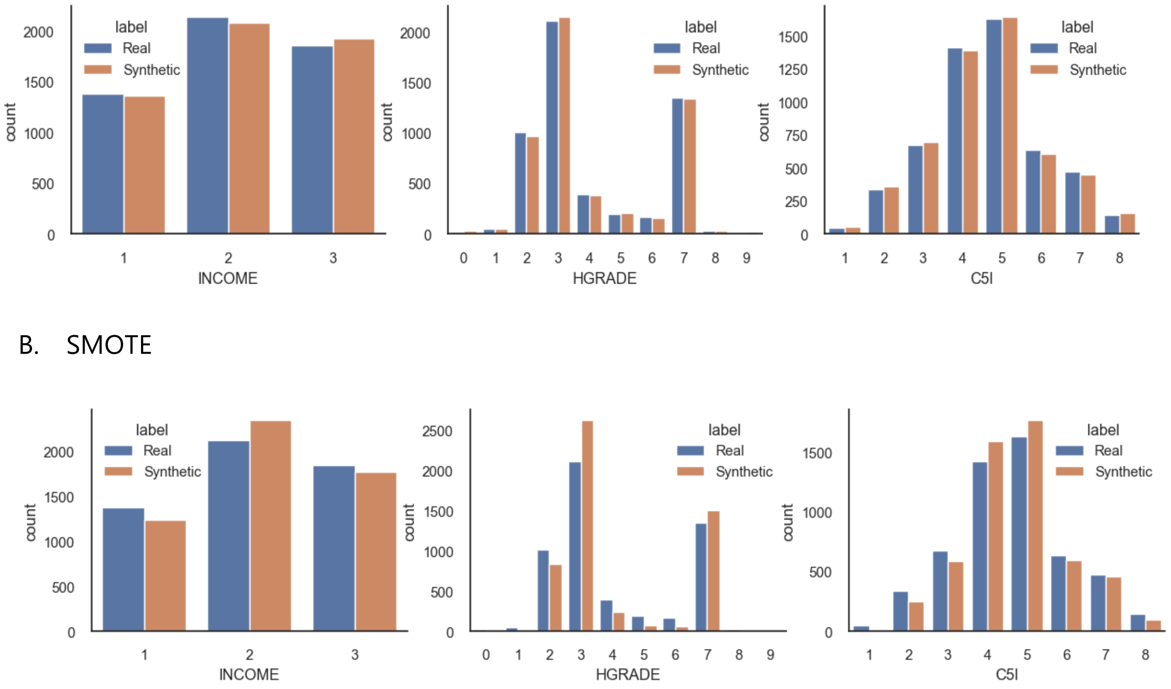
In assessing the quality of the generated synthetic data in terms of similarity in statistical properties, we calculated the statistical similarity score as the average similarity in the histogram and correlation heatmap.

Histogram Plots

We analyzed a total of 36 variables, which sample histogram plots are shown in Figure 2. The GC model best captured the statistical properties of the original dataset. On the other hand, SMOTE had the lowest similarity score among the tested algorithms. The full results are shown in Appendix B.

Histogram of Real and Synthetic Data

Figure 2



A. Gaussian Copula (GC)

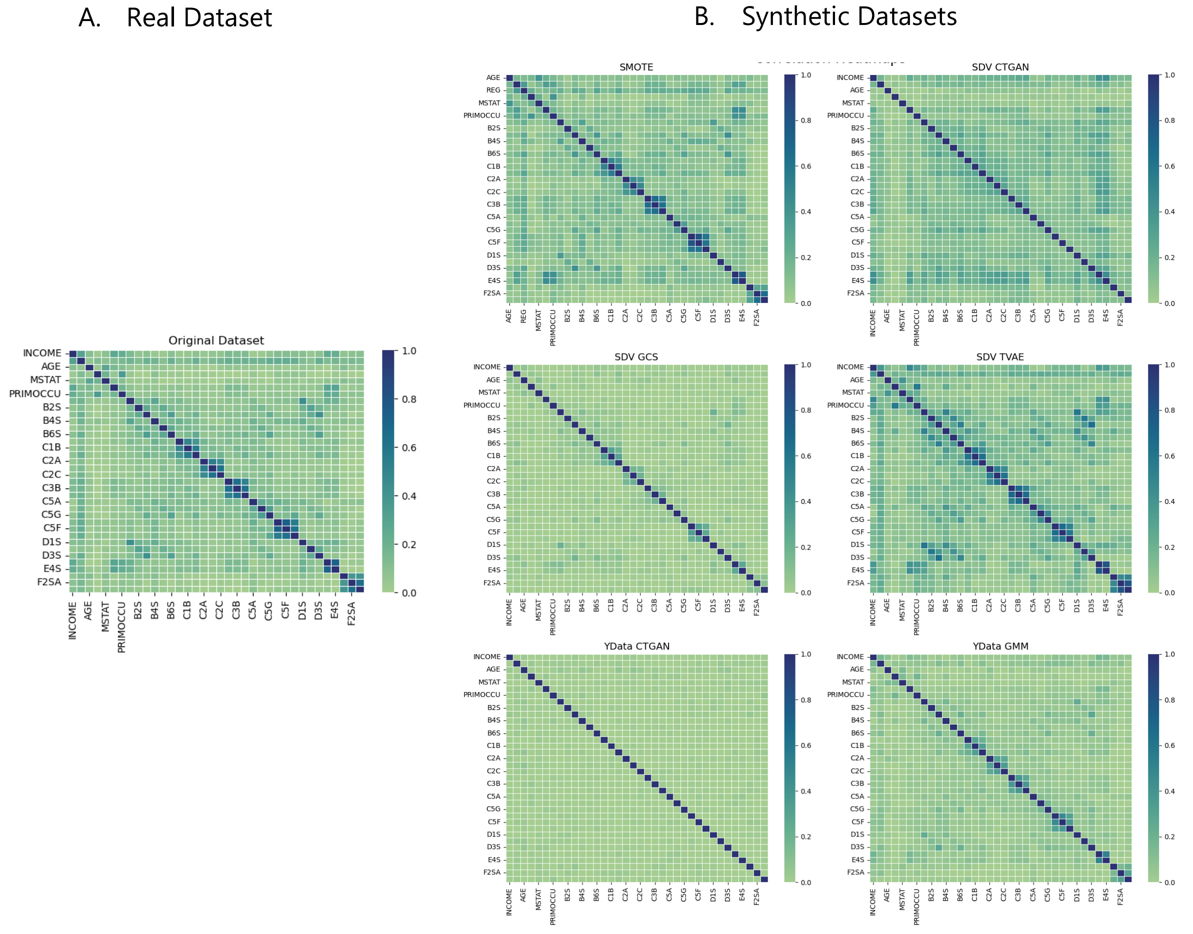
B. SMOTE

Correlation heatmaps

In addition to the generated histogram plots, we also evaluated the correlation between variables. Figure 3 shows the heatmaps of Cramer’s V correlation for the original and synthetic datasets. The heatmaps of SMOTE and SDV TVAE synthetic datasets highly resemble the original dataset, given that these captured specific portions of highly correlated variables (in dark blue). Although slightly faint, the YData GMM and SDV GC could identify correlated variables correctly.

Correlation Heatmaps of Real and Synthetic Datasets

Figure 3



A. Real Dataset

B. Synthetic Datasets

Statistical Similarity Score

We computed the similarity scores to support the histogram plots and the Cramer’s V values heatmap for the 36 variables, , as shown in Table 3. Each row in the table shows the similarity of the values generated from the real and the corresponding synthetic data. Then, we computed the mean of the Bray-Curtis and Cosine Similarity scores to derive an overall *statistical similarity score*.

Statistical Similarity Score[1]

Values in parentheses are the standard deviation of the computed scores. Table 3

Name of Python Library	Algorithm	Histogram Similarity	Correlation Similarity	Overall Statistical Similarity Score
YData	GMM	0.9696 (±0.0247)	0.9306	0.9501
SDV	TVAE	0.9381 (±0.0427)	0.9171	0.9276
In-house	SMOTE	0.8882 (± 0.0599)	0.9666	0.9274
SDV	GC	0.9893 (±0.0062)	0.8377	0.9135
SDV	CTGANs	0.8884 (±0.0699)	0.7226	0.8055
YData	CTGANs	0.8974 (±0.0621)	0.4023	0.6499

[1] Statistical Similarity Score is the average of the computed histogram and correlation similarity scores.

Sources: Authors’ computations

Most algorithms produced a statistical similarity score above 0.9 except for CTGANs, indicating poor data fidelity and inability to resemble the original data’s statistical properties.

In contrast, for histogram similarity, YData GMM datasets yielded the highest similarity score (0.9893), showing that the algorithm best preserved the count distribution of the variables. In contrast, SMOTE and CTGANs presented the lowest histogram similarity, coinciding with the histogram plots.

Meanwhile, SMOTE had the highest correlation similarity (0.9666), corroborating its high cosine similarity scores heatmap vis-à-vis the original.

For the combined statistical similarity, GMM and TVAE presented the most accurate fully synthetic data compared to the rest of the algorithms with scores of 0.9501 and 0.9276. Closely following is SMOTE. However, it presented a low histogram similarity score, making it unfit for exploratory data analysis. On the other hand, GC datasets preserved the histogram well, but the correlation of variables was farther than the actual dataset, with a cosine similarity score of 0.8377.

Data Utility

We measured the usability of the synthetic dataset by evaluating the performance of each algorithm compared to the original data through these metrics: accuracy, AUC, recall, precision, F1 score, Kappa, and the Matthews Correlation Coefficient (MCC)[2]. We also evaluated the Feature Importance to supplement the machine learning results.

In Table 4, we compared the performance of ML models trained on both synthetic and actual datasets using a Gradient Boosting Classifier (GBC) since this gave the best performance compared to other ML models (e.g., naïve Bayes, k-neighbors, logistic regression), and validated on test data from the actual dataset.

SMOTE, TVAE, and GMM performed well when compared with the actual dataset. Meanwhile, GC- and GAN-based synthetic datasets provided significantly worse performance on all metrics and, thus, is not recommended for research purposes.

Xu et al. (2020) generated synthetic tabular data using GANs and found that TVAE outperforms GANs in classification and regression tasks. However, GANs still offered several favorable attributes, making it easier to learn data distributions better than other models, such as Bayesian networks.

Machine Learning Performance of Synthetic Datasets[1]

Utility is evaluated based on the AUC score Table 4

Dataset/Python Library	Algorithm	Accuracy	AUC	Recall	Precision	F1	Kappa	MCC
Actual Data		0.7852	0.9493	0.7852	0.7913	0.786	0.7303	0.7313
Synthetic Data:								
	SMOTE	0.7459	0.9259	0.7459	0.751	0.7456	0.6815	0.6828
SDV	TVAE	0.7247	0.9217	0.7247	0.7255	0.7239	0.6529	0.6531
YData	GMM	0.7171	0.9039	0.7171	0.7125	0.7123	0.6411	0.6419
SDV	GC	0.4276	0.739	0.4276	0.3794	0.3844	0.2539	0.2593
SDV	CTGAN	0.438	0.7121	0.438	0.4174	0.3901	0.2681	0.2846
YData	CTGAN	0.2686	0.5433	0.2686	0.2347	0.2271	0.0071	0.0076

[1] The ML performance is evaluated based on the difference between the actual and synthetic data performances. The lower the difference, the better the performance.

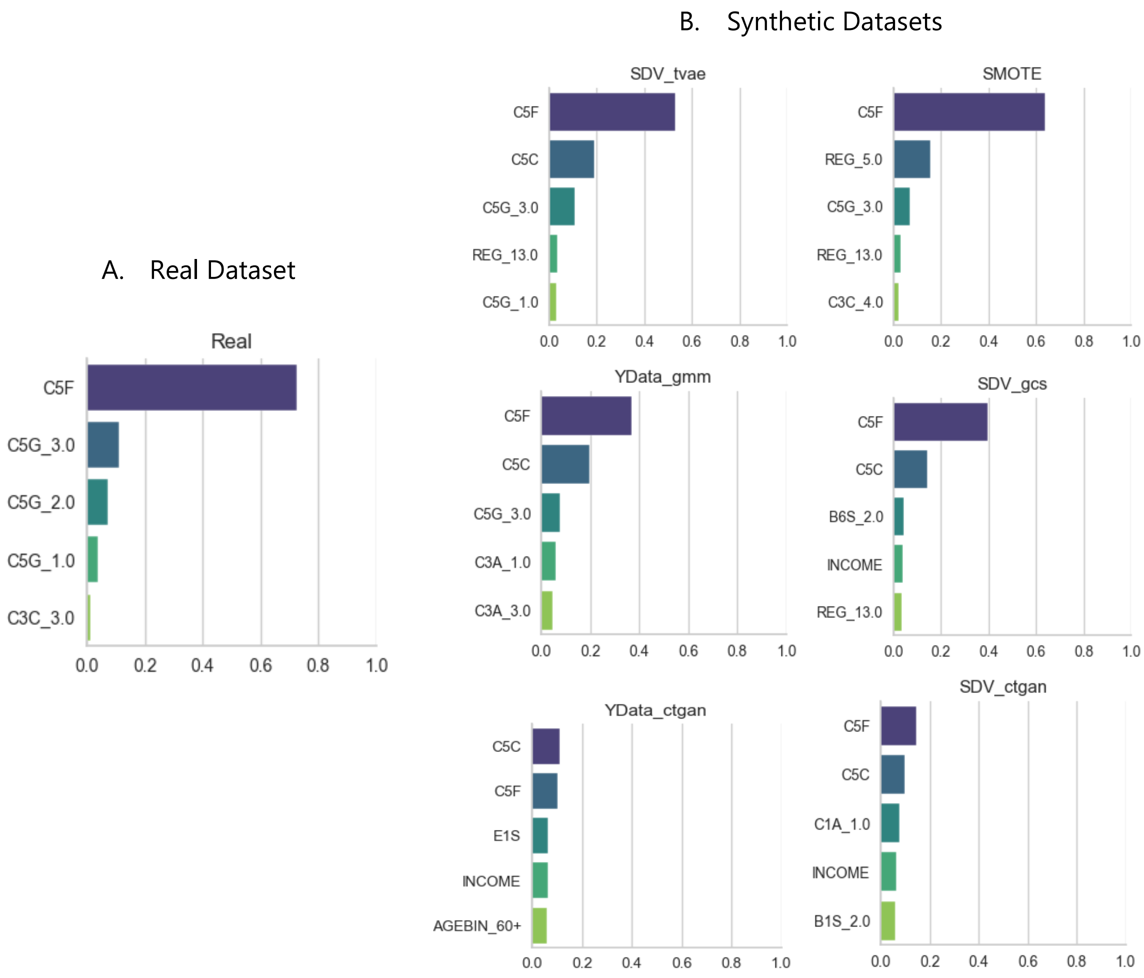
Sources: Authors' computations

Furthermore, the feature importance scores helped interpret machine learning performance by determining the relative significance of each feature within the model. This capability allowed us to pinpoint key variables and enhance our understanding of the problem. We condisder the utility of synthetic data to be high when it successfully preserves both the order and magnitude of feature importance.

[2] The metrics are based on the default PyCaret machine learning performance evaluation. Kappa or Cohen's Kappa is an evaluation metric quantifying the level of agreement between two or more raters (or models) in the classification of categorical data, while considering the agreement between two or more raters. The Matthews Correlation Coefficient or MCC is an evaluation metric measuring the quality of predictions by considering both true positive and true negative results while accounting for the balance between classes.

Feature Importance Scores of Real and Synthetic Datasets

Figure 4



A. Real Dataset

B. Synthetic Datasets

TVAE, GMM, and SMOTE matched the top predictor variables (C5F and C5G) of the ML model developed using the actual dataset, aligning with the observed ML performance in Table 4.

In contrast, we found that GC and CTGANs exhibited notably distinct feature distribution patterns compared to the actual and other synthetic datasets examined in this study.

Data Privacy and Disclosure Risk

The membership inference score gauges the susceptibility of individual data points to membership inference attacks. Even without access to the actual dataset, attackers can reveal the data used to create the synthetic data, posing a risk of re-identification and privacy breaches.

Here, we compared three metrics to assess the privacy of synthetic data: accuracy, AUC score, and precision. A low score implies an increased risk of inference, compromising individual record privacy. In contrast, a high score suggests that an attacker is unlikely to determine if a record was part of the original dataset. Table 5 shows the membership inference scores.

Membership Inference Score[1]

Proxy for Privacy score

Table 5

Name of Python Library	Algorithm	Accuracy	AUC Score	Precision
YData	CTGANs	0.9660	0.9972	0.9754
SDV	CTGANs	0.9451	0.9929	0.9574
SDV	GC	0.9148	0.9796	0.9260
SDV	TVAE	0.8181	0.9246	0.8472
YData	GMM	0.7911	0.9032	0.8199
	SMOTE	0.7664	0.7784	0.7134

[1] The privacy score is proxied by the concept of membership inference score, which relies on precision as the primary measure.

Sources: Author’s computations

We found that GAN-based models consistently achieved higher scores in this metric, providing compelling evidence that attempts to reverse engineer the model for synthetic dataset generation and reveal the actual records are highly improbable with such models.

However, SMOTE is an exception to this trend. By design, SMOTE generated synthetic data more systematically and straightforwardly, making it relatively easier to replicate the underlying process. Consequently, SMOTE-generated synthetic data were more vulnerable to attacks for exposing actual data points.

It is worth noting that the significant gap between the AUC score and the precision score reveals the impact of false positives during the classification process. Such disparity underscored the importance of minimizing false positives to enhance the overall effectiveness of privacy-preserving measures in ML models.

Evaluation of Best Synthetic Dataset

We devised a straightforward scoring system to comprehensively assess and identify the most suitable synthetic dataset, as discussed in Table 2. We derived these scores from the algorithms’ performance across the evaluation metrics discussed in earlier sections.

Synthetic Data Evaluation Matrix[1]

3 – Excellent, 2 – Fair, 1 - Poor

Table 6

Metric	Algorithm	Fidelity	Utility	Privacy	Overall
YData	CTGANs	1	1	3	1.7
SDV	CTGANs	1	1	3	1.7
SDV	GC	2	1	3	2.0
SDV	TVAE	3	2	2	2.3
YData	GMM	3	2	2	2.3
In-house	SMOTE	2	3	1	2.0

[1] The evaluation matrix is based on the following resource: asq.org/quality-resources/decision-matrix

Sources: Authors' computations

Based on this scoring system, TVAE and GMM emerged as the top-performing synthetic data generation models, excelling in data fidelity, utility, and privacy, with an average score of 2.3. Meanwhile, GC and SMOTE ranked second with average scores of 2.0. CTGAN for YData and SDV ranked last, consistently scoring poorly across all evaluation metrics except privacy.

High-quality synthetic datasets must be based on an algorithm that performs exceptionally well in all three aspects. Of the six synthetic data generation algorithms tested, only GMM and TVAE algorithms have shown higher-quality synthetic CES data for research. On the other hand, CTGAN-based synthetic data ranked highest in data privacy. Still, CTGAN could not capture the inherent properties of the actual dataset, agreeing with the results of Little et al. (2021) and Pathare et al. (2023).

6. Conclusion and Recommendations

Synthetic data generation is valuable for addressing data-sharing concerns with external entities, especially researchers outside the BSP. In harnessing the potential of synthetic data for research and data-driven decision-making, it is imperative to remain prudent about preserving privacy, maintaining data quality, and adhering to ethical and regulatory standards. By maximizing the potential of synthetic datasets, data-sharing protocols can be streamlined through automated generation and evaluation processes while upholding data privacy and confidentiality.

We explored the generation of synthetic datasets and presented a synthetic tabular data generation pipeline using the CES dataset, primarily composed of categorical variables. Unlike previous studies using cleaned datasets from public repositories (e.g., UCI Machine Learning repository), we utilized a raw and unprocessed dataset, adding complexity in maintaining data integrity during cleaning and pre-processing before employing machine learning algorithms for synthetic data generation.

The algorithms we used in this study encompass traditional statistical methods (e.g., SMOTE, GMM, and GC) and cutting-edge deep learning techniques (e.g., TVAE and GANs) through open-source libraries dedicated to synthetic data generation, specifically YData Synthetic and SDV. Both methodologies offered ease of use and flexibility, allowing control over data processing parameters to enhance model performance and synthetic data quality.

We found that GMM and TVAE are the most effective algorithms for generating synthetic data from the CES dataset, meeting evaluation criteria for data fidelity, utility, and privacy. Notably, GAN-based algorithms excelled in preserving data privacy but not the statistical properties of the actual dataset.

Future research endeavors may explore synthetic data generation for survey datasets containing various data types (i.e., numerical and a combination of categorical and numerical data), such as the BSP's Consumer Finance Survey. Furthermore, central banks often use time-series data for macroeconomic research. Future studies may explore creating synthetic time-series data to extend short datasets by adding extra data points. Lastly, future works may delve into developing a more systematic approach to the synthetic data generation process. It is worth emphasizing that this study serves as proof of concept and sets the stage for operationalizing the synthetic data generation pipeline in the BSP.

References

Assefa, S. A., Dervovic, D., Mahfouz, M., Tillman, R. E., Reddy, P., & Veloso, M. (2020). Generating synthetic data in Finance. Proceedings of the First ACM International Conference on AI in Finance. https://doi.org/10.1145/3383455.3422554

Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., & Tramer, F. (2022). Membership inference attacks from first principles. 2022 IEEE Symposium on Security and Privacy (SP). https://doi.org/10.1109/sp46214.2022.9833649

Cramér, H. (1946). Mathematical Methods of Statistics (PMS-9). In *Princeton University Press eBooks*. https://doi.org/10.1515/9781400883868

Irving Fisher Committee on Central Bank Statistics. (2015). (rep.). Data-sharing: issues and good practices. Retrieved 2023, from https://www.bis.org/ifc/events/7ifc-tf-report-datasharing.pdf.

Heyburn, R., Bond, R., Black, M., Mulvenna, M., Wallace, J., Rankin, D., & Cleland, B. (2018). Machine learning using synthetic and real data: Similarity of evaluation metrics for different healthcare datasets and for different algorithms. *Data Science and Knowledge Engineering for Sensing Decision Support*. https://doi.org/10.1142/9789813273238_0160

Koblents, E. & Megia, A.L. (2023). "Joint secondary anonymisation of categorical and numerical variables in sensitive time series microdata - novel approach for Statistical Disclosure Control of a sensitive microdata set published in BE," IFC Bulletins chapters, in: Bank for International Settlements (ed.), Post-pandemic landscape for central bank statistics, volume 58, Bank for International Settlements.

Little, C., Elliot, M., Allmendinger, R., & Samani, S. S. (2021). Generative Adversarial Networks for Synthetic Data Generation: A Comparative study. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2112.01925

Mukherjee, M., & Khushi, M. (2021). SMOTE-ENC: a novel SMOTE-Based method to generate synthetic data for nominal and continuous features. *Applied System Innovation*, *4*(1), 18. https://doi.org/10.3390/asi4010018

Pan, J., Pham, V., Dorairaj, M., Chen, H., & Lee, J. (2020). Adversarial Validation approach to concept drift problem in user targeting automation systems at Uber. arXiv (Cornell University). https://arxiv.org/pdf/2004.03045.pdf

Pathare, A., Mangrulkar, R. S., 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. *IEEE*. https://doi.org/10.1109/dsaa.2016.49

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). SciKit-Learn: Machine Learning in Python. *HAL (Le Centre Pour La Communication Scientifique Directe)*. https://hal.inria.fr/hal-00650905

Raghunathan, T. E. (2021). Synthetic data. *Annual Review of Statistics and Its Application*, *8*(1), 129–140. https://doi.org/10.1146/annurev-statistics-040720-031848

Platzer, M., & Reutterer, T. (2021). Holdout-based empirical assessment of mixed-type synthetic data. *Frontiers in Big Data*, *4*. https://doi.org/10.3389/fdata.2021.679939

Sallier, K. (2020). Toward more user-centric data access solutions: Producing synthetic data of high analytical value by data synthesis1. *Statistical Journal of the IAOS*. https://doi.org/10.3233/sji-200682

United Nations Economic Commission for Europe. (2022). *Synthetic Data for Official Statistics: A Starter Guide*. https://unece.org/sites/default/files/2022-11/ECECESSTAT20226.pdf

Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling Tabular data using Conditional GAN. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1907.00503

Zhang, Y., Zaidi, N. A., Zhou, J., & Li, G. (2023). Interpretable tabular data generation. *Knowledge and Information Systems*, *65*(7), 2935–2963. https://doi.org/10.1007/s10115-023-01834-5

Zhao, Z., Kunar, A., Van Der Scheer, H., Birke, R., & Chen, L. Y. (2021). CTAB-GAN: Effective table data Synthesizing. *arXiv (Cornell University)*. https://arxiv.org/pdf/2102.08369.pdf

Appendix A

List of Variables

Variable Name	Description	Values
INCOME	Income Class	1 – Low-income 2 – Middle-income 3 – High-income
REG	Region	1 – Region 1 – Ilocos 2 – Region 2 – Cagayan Valley 3 – Region 3 – Central Luzon 4 – Region 4 – CALABARZON 5 – Region 5 – Bicol 6 – Region 6 – Western Visayas 7 – Region 7 – Central Visayas 8 – Region 8 – Eastern Visayas 9 – Region 9 – Western Mindanao 10 – Region 10 – Northern Mindanao 11 – Region 11 – Southern Mindanao 12 – Region 12 – Central Mindanao 13 – Region 13 – National Capital Region 14 – Region 14 – Cordillera Administrative Region 15 – Region 15 – Autonomous Region in Muslim Mindanao 16 – Region 16 – Caraga 17 – MIMAROPA
AGE	Age	None
SEX	Sex	1 – Male 2 – Female
MSTAT	Marriage Status	1 – Single 2 – Married 3 – Common-law/Live-in 4 – Widowed 5 – Divorced 6 – Separated 7 – Annulled 8 - Unknown
HGRADE	Highest Educational Attainment	0 - No Grade Completed 1 - Early Childhood Education 2 - Primary Education 3 - Lower Secondary Education 4 - Upper Secondary Education 5 - Post-Secondary Non-Tertiary Education 6 - Short-Cycle Tertiary Education 7 - Bachelor Level Education or Equivalent 8 - Master Level Education or Equivalent 9 - Doctor Level Education or Equivalent

List of Variables (cont.)

Variable Name	Description	Values
PRIMOCCU	Primary Occupation	0 - Armed Forces Occupations 1 - Managers 2 - Professionals 3 - Technicians and Associate Professionals 4 - Clerical support workers 5 - Service & sales workers 6 - Skilled agricultural, forestry & fishery workers 7 - Craft & related trades workers 8 - Plant & Machine Operators & Assemblers 9 - Elementary Occupations
B1S	Present Financial Situation	1 - Better 2 - Same 3 - Worse
B2S	Financial Situation after 3 months	1 - Better 2 - Same 3 - Worse
B3S	Financial Situation after 12 months	1 - Better 2 - Same 3 - Worse
B4S	Present Economic Condition of the Country	1 - Better 2 - Same 3 - Worse
B5S	Economic Condition of the Country after 3 Months	1 - Better 2 - Same 3 - Worse
B6S	Economic Condition of the Country after 12 Months	1 - Better 2 - Same 3 - Worse
C1A	Number of Unemployed Persons for the Current Quarter	1 - Increase 2 - Same as this period 3 - Decrease
C1B	Number of Unemployed Persons for the Next Quarter	1 - Increase 2 - Same as this period 3 - Decrease
C1C	Number of Unemployed Persons for the Next 12 Months	1 - Increase 2 - Same as this period 3 - Decrease
C2A	Level of Interest Rates for Borrowing Money for the Current Quarter	1 - Increase 2 - Same as this period 3 - Decrease
C2B	Level of Interest Rates for Borrowing Money for the Next Quarter	1 - Increase 2 - Same as this period 3 - Decrease

List of Variables (cont.)

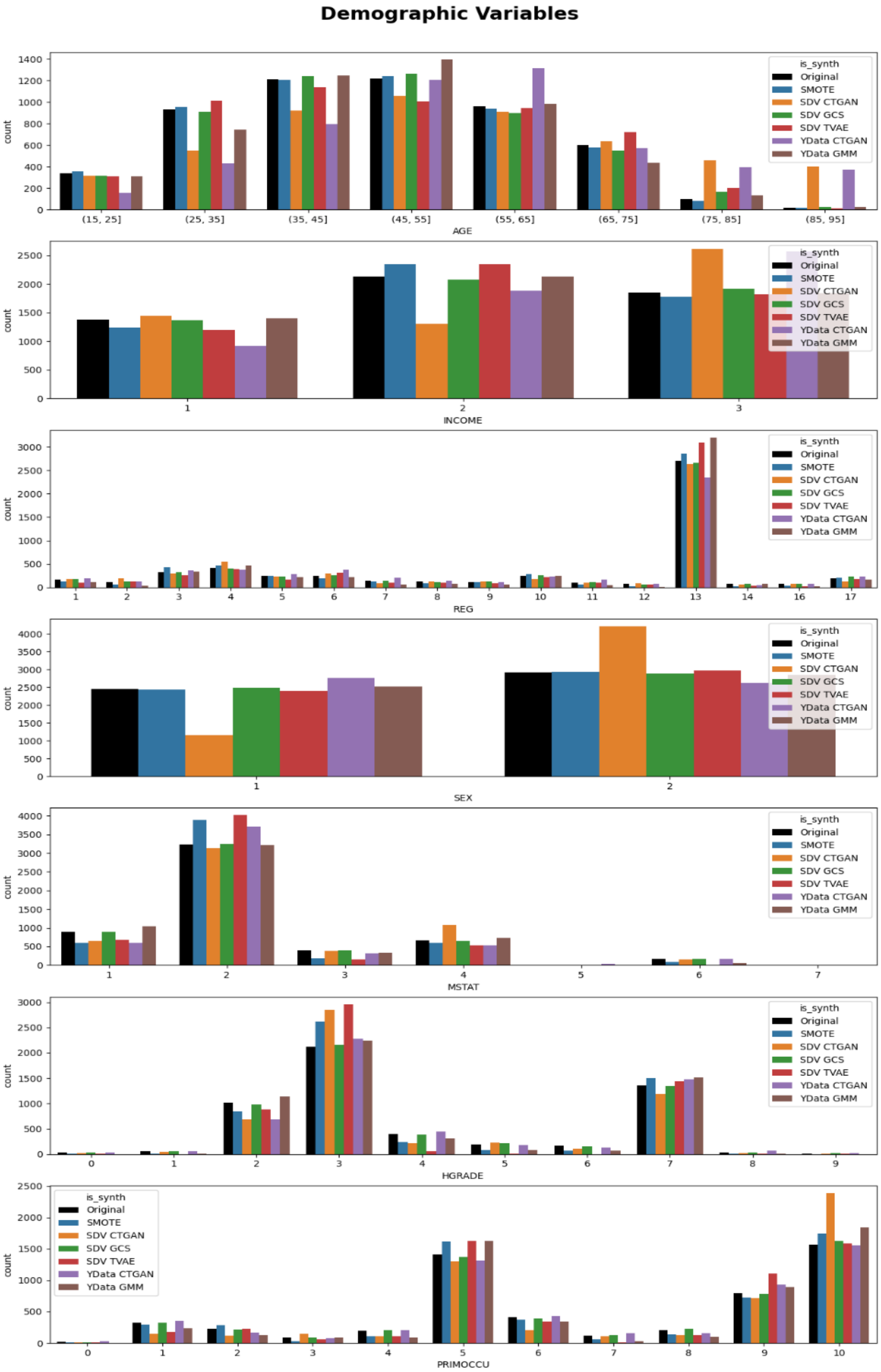
Variable Name	Description	Values
C2C	Level of Interest Rates for Borrowing Money for the Next 12 Months	1 - Increase 2 - Same as this period 3 - Decrease
C3A	Peso-Dollar Exchange Rate for the Current Quarter	1 - Appreciate 2 - Same as this period 3 - Depreciate 4 - Don't know
C3B	Peso-Dollar Exchange Rate for the Next Quarter	1 - Appreciate 2 - Same as this period 3 - Depreciate 4 - Don't know
C3C	Peso-Dollar Exchange Rate for the Next 12 Months	1 - Appreciate 2 - Same as this period 3 - Depreciate 4 - Don't know
C5A	Inflation rate will rise, remain, unchanged or fall in the current quarter	1 - Will go up 2 - Remain unchanged 3 - Will go down
C5C	Inflation rate for the current quarter	1 - Less than 0% 2 - Equal to 0% 3 - 0.1% to 1.9% 4 - 2% to 3.9% 5 - 4% to 5.9% 6 - 6% to 7.9% 7 - 8% to 9.9% 8 - 10% or more
C5D	Inflation rate will rise, remain, unchanged or fall in the next quarter	1 - Will go up 2 - Remain unchanged 3 - Will go down
C5F	Inflation rate for the current year	1 - Less than 0% 2 - Equal to 0% 3 - 0.1% to 1.9% 4 - 2% to 3.9% 5 - 4% to 5.9% 6 - 6% to 7.9% 7 - 8% to 9.9% 8 - 10% or more
C5G	Inflation rate will rise, remain, unchanged or fall in the next quarter	1 - Will go up 2 - Remain unchanged 3 - Will go down

List of Variables (cont.)

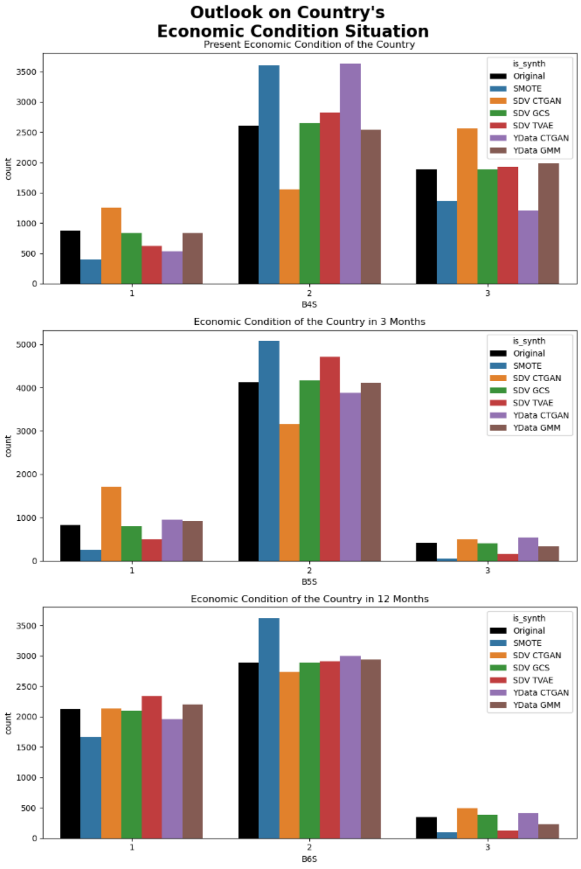
Variable Name	Description	Values
C5I	Inflation rate for the next 12 months	1 - Less than 0% 2 - Equal to 0% 3 - 0.1% to 1.9% 4 - 2% to 3.9% 5 - 4% to 5.9% 6 - 6% to 7.9% 7 - 8% to 9.9% 8 - 10% or more
D1S	Current Level Family Income	1 - Went up 2 - Same as now 3 - Went down
D2S	Family Income Next 3 Months	1 - Will go up 2 - Same as now 3 - Will go down
D3S	Family Income Next 12 Months	1 - Will go up 2 - Same as now 3 - Will go down
E1S	Has family savings	1 - Yes 2 - No
E4S	The family could set aside money for savings during the current quarter	1 - Yes 2 - No
F1SA	Outstanding loan	1 - Yes 2 - No
F2SA	Plan to apply loan in the next quarter	1 - Yes 2 - No
F3SA	Plan to apply loan in the next 12 months	1 - Yes 2 - No

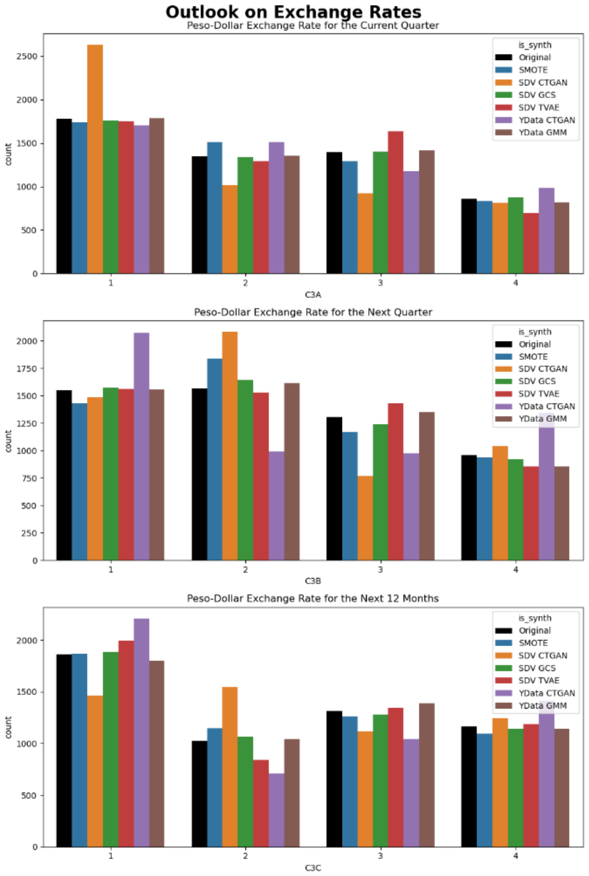
Appendix B

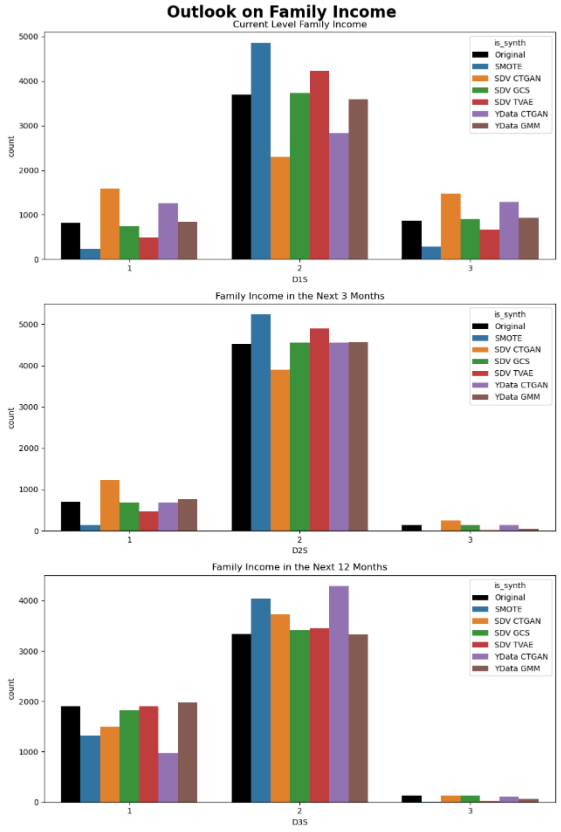
Comparison of histogram plot for each variable for actual vs. synthetic data

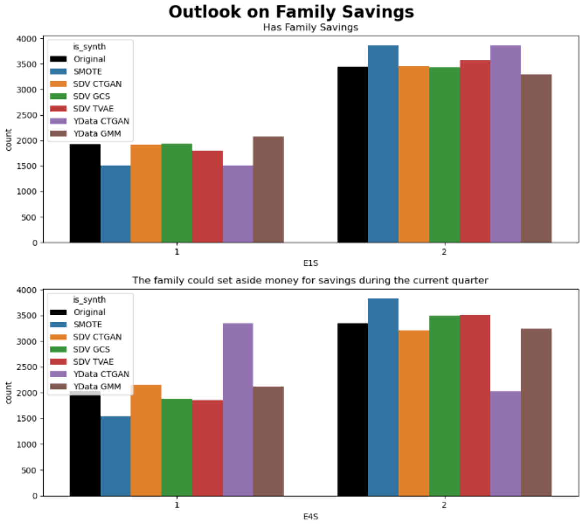


Comparison of histogram plot for actual vs. synthetic data (cont.)

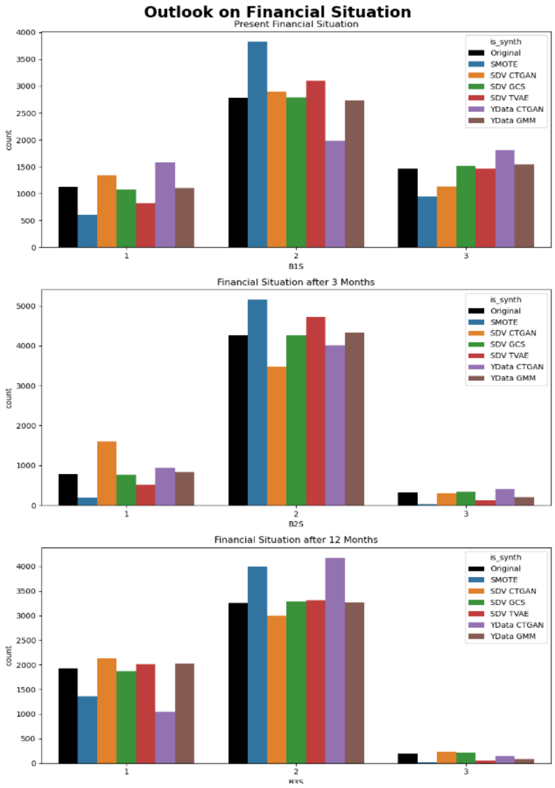


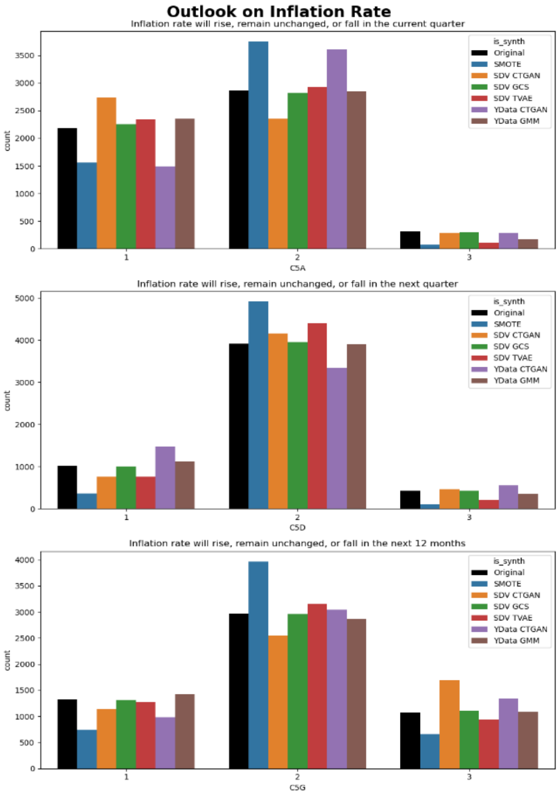


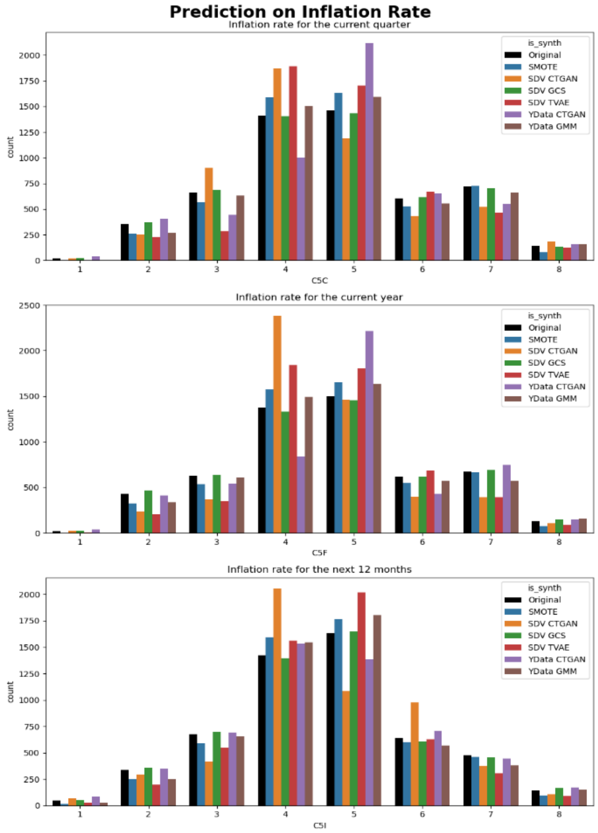


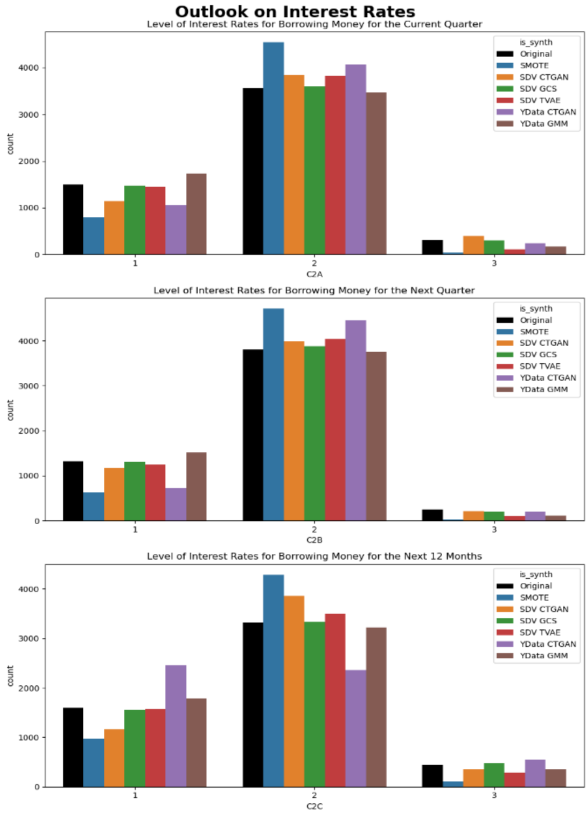


Comparison of histogram plot for actual vs. synthetic data (cont.)

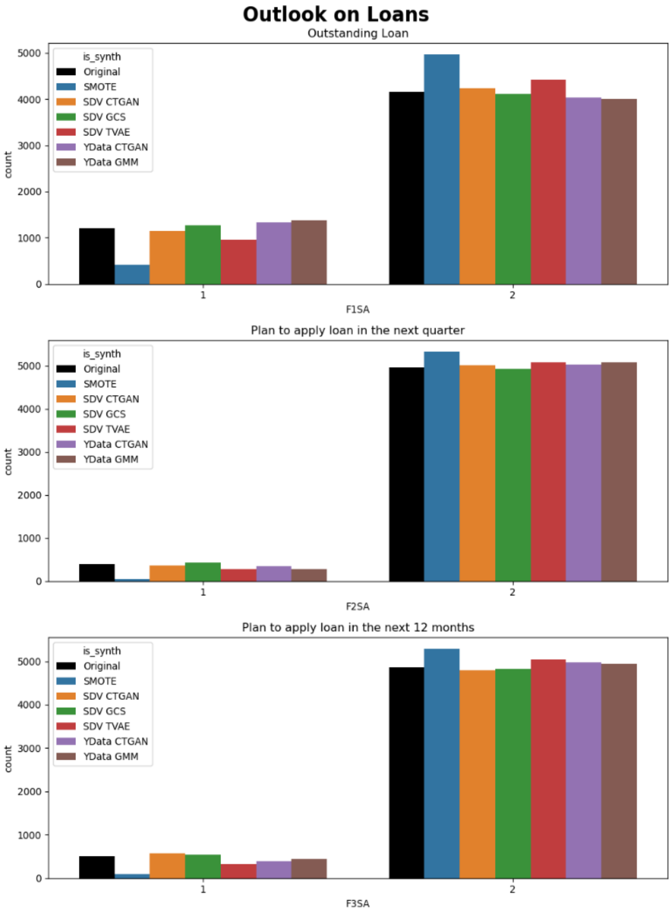


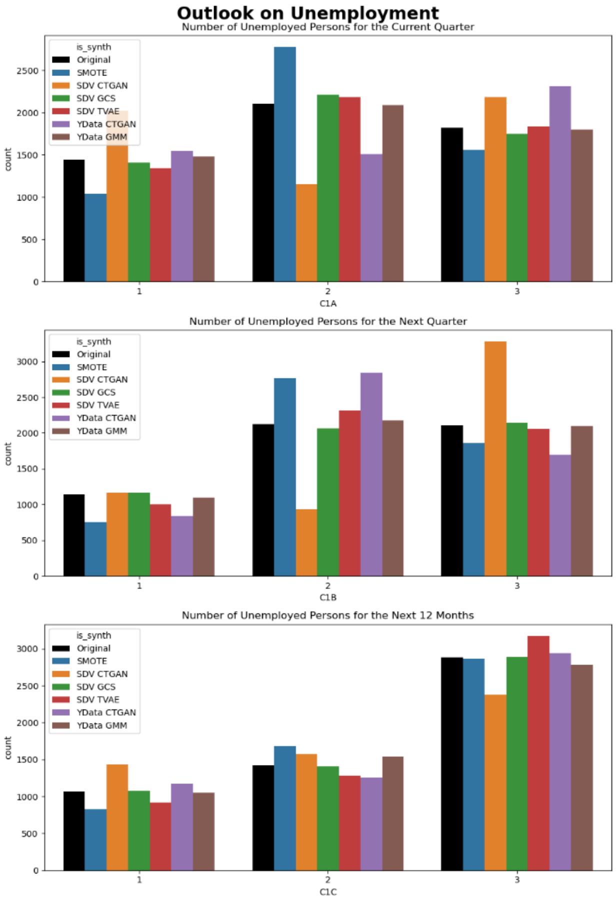






Comparison of histogram plot for actual vs. synthetic data (cont.)





Research for All: Exploring machine learning applications in generating synthetic datasets

CARMELITA ESCLANDA-LO
GABRIEL MASANGKAY
CHELSEA ANNE ONG
ROSSVERN REYES

3RD IFC AND BANK OF ITALY WORKSHOP ON "DATA SCIENCE IN CENTRAL BANKING: ENHANCING THE ACCESS TO AND SHARING OF DATA", OCTOBER 17-19, 2023

* The views expressed herein are those of the authors' only and do not necessarily reflect those of the Bangko Sentral ng Pilipinas

OUTLINE

01 MOTIVATION

02 RELATED STUDIES

03 DATA

04 METHODOLOGY

05 RESULTS

06 KEY TAKEWAYS AND RECOMMENDATIONS

Motivation and Objectives

Ease data sharing procedures

Explore AI for synthetic data generation

Generate quality and private data for research use

Related Studies

Data Sharing Practices

Data sharing frameworks are in place and delegated entities enforce these frameworks.

The BSP has developed the **Data Governance Manual** which specifies sharing data to external parties and protecting sensitive information.

Meanwhile, **National Statistics Offices around the world have developed and operationalized synthetic datasets** for public data dissemination, improve efficiency of data sharing processes.

Use of Synthetic Data for Research

For methodologies on generation of synthetic tabular data, most studies explore **Generative Adversarial Networks (GANs) and Tree-based models.**

"As the utility of synthetic data increases, the disclosure risk increases exponentially."

Classification: GENERAL

Data

CONSUMER EXPECTATIONS SURVEY

Quarterly survey conducted by BSP to gather information from Filipino households regarding **sentiments on various economic indicators.**

SAMPLE VARIABLES

	Description	*Values*
Identifier Variables		
AGE	Age	0-100
INCOME	Income Group	Low, Middle, High
SEX	Sex	Male, Female
....	....	....
Response Variables		
C5C	Inflation Rate in the Current Quarter	Less than 0%, 0.1%-1.9%, ……
E1S	Has Family Savings	Yes, No
B1S	Present Financial Situation	Better, Same, Worse
....	....	....

Methodology

Data Collection and Preprocessing

Processing is done to preselect columns, **address missing data, differing data types**. Options for **variable selection, data binning, partial synthesis** are covered in this study.

Generate Synthetic Data

The following algorithms will be tested using in-house and **open-source packages** (e.g., Synthetic Data Vault (SDV), YData Synthetic):

- **SMOTE**
- **Gaussian Mixture Models (GMM)**
- **Gaussian Copula (GC)**
- **Tabular Variational Autoencoders (TVAE)**
- **Conditional Tabular Generative Adversarial Networks (CTGAN)**

Evaluate Synthetic Data

Assess whether synthetic datasets can be used as an alternative dataset. These shall be evaluated based on three key dimensions: **fidelity, utility, privacy.**

Synthetic Data Evaluation

Assess whether **synthetic dataset can be used as an alternative dataset for research use.**

Data Fidelity

- Statistical Similarity
  - Histogram
  - Correlation

Data Utility

- Machine Learning Performance
  - Accuracy
  - AUC
  - Recall
  - Precision
  - F1
  - Kappa
  - MCC

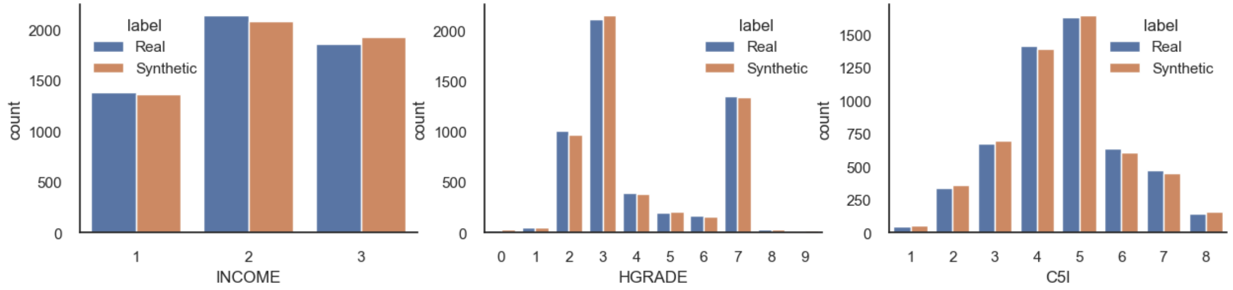
Data Privacy

- Membership Inference
  - Accuracy
  - AUC
  - Precision

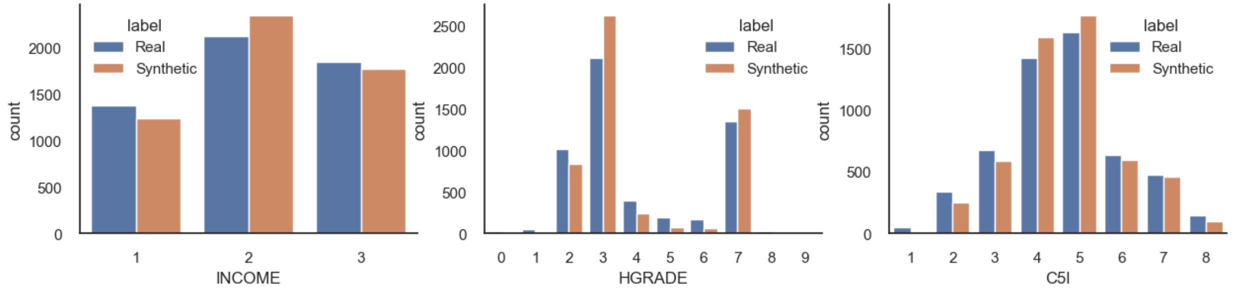
Data Fidelity

HISTOGRAM COUNTPLOTS

A total of 36 variables is analyzed to **compare the count distribution** for real and synthetic datasets.



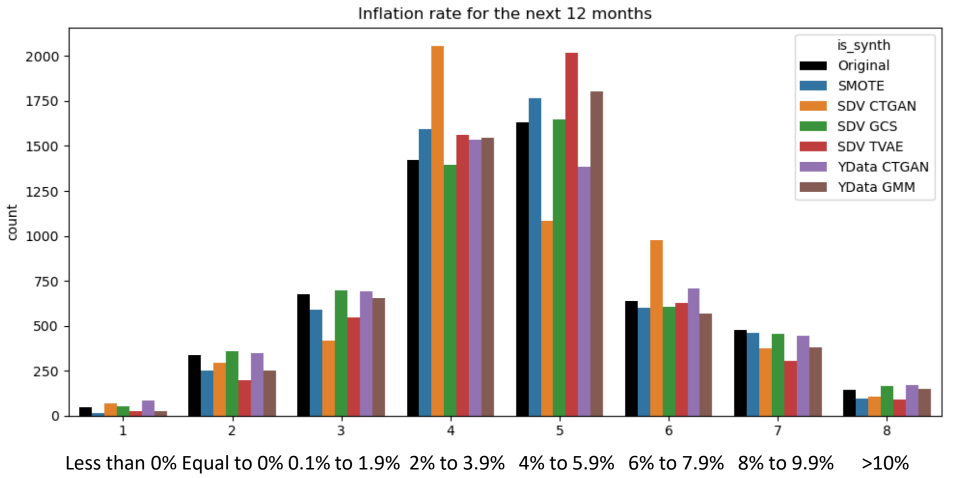
Gaussian Copula
Similarity score = 0.9893 (± 0.0062)



SMOTE
Similarity score = 0.8882 (± 0.0599)

**Values in parentheses are standard deviations*

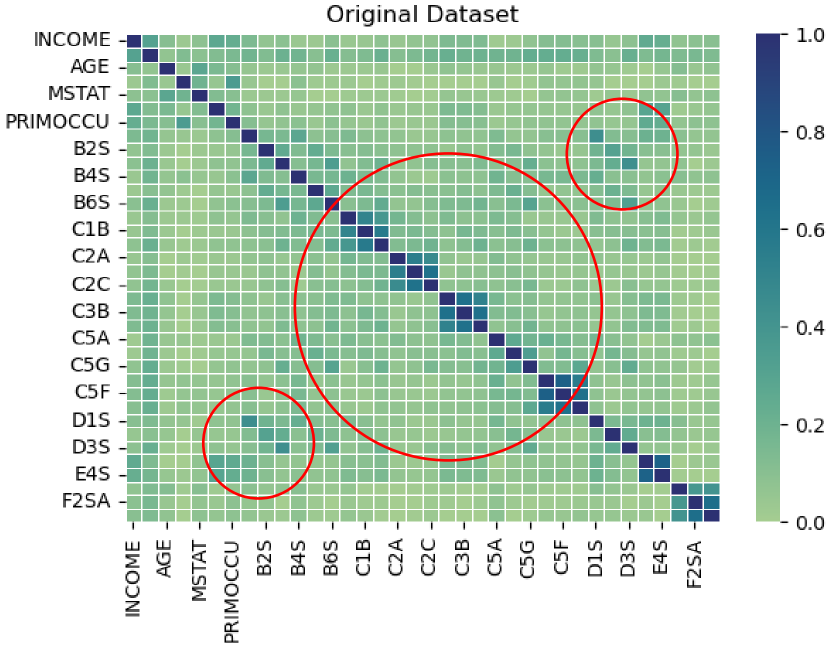
TARGET VARIABLE (C5I)

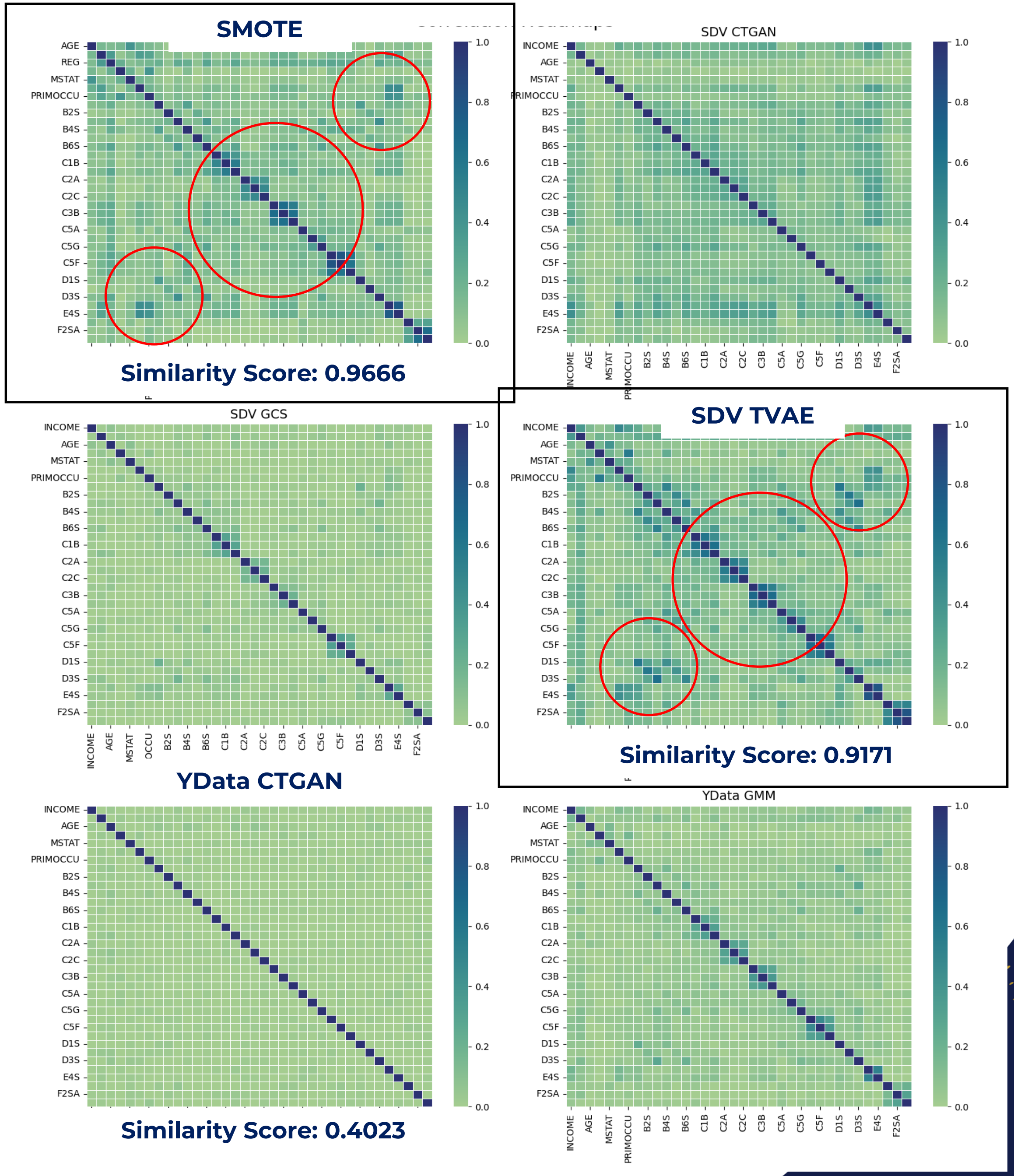


Data Fidelity

CORRELATION HEATMAP – CRAMER'S V

Cramér's V is used to determine whether a **significant relationship exists between two categorical variables.**





Similarity Score: 0.9666

Similarity Score: 0.9171

Similarity Score: 0.4023

Data Fidelity

STATISTICAL SIMILARITY

The Statistical Similarity score is computed as the **average of the histogram and correlation similarity scores.**

Python Library	Algorithm	Bray-Curtis Similarity Scores (Histogram)	Cosine Similarity Scores (Correlation)	Statistical Similarity Score
YData	GMM	0.9696 (± 0.0247)	0.9306	0.9501
SDV	TVAE	0.9381 (± 0.0427)	0.9171	0.9276
In-house	SMOTE	0.8882 (± 0.0599)	**0.9666**	0.9274
SDV	GC	**0.9893 (± 0.0062)**	0.8377	0.9135
SDV	CTGANs	0.8884 (± 0.0699)	0.7226	0.8055
YData	CTGANs	0.8974 (± 0.0621)	0.4023	0.6499

**Values in parentheses are standard deviations* is italic: **Values in parentheses are standard deviations*

Data Utility

MACHINE LEARNING PERFORMANCE

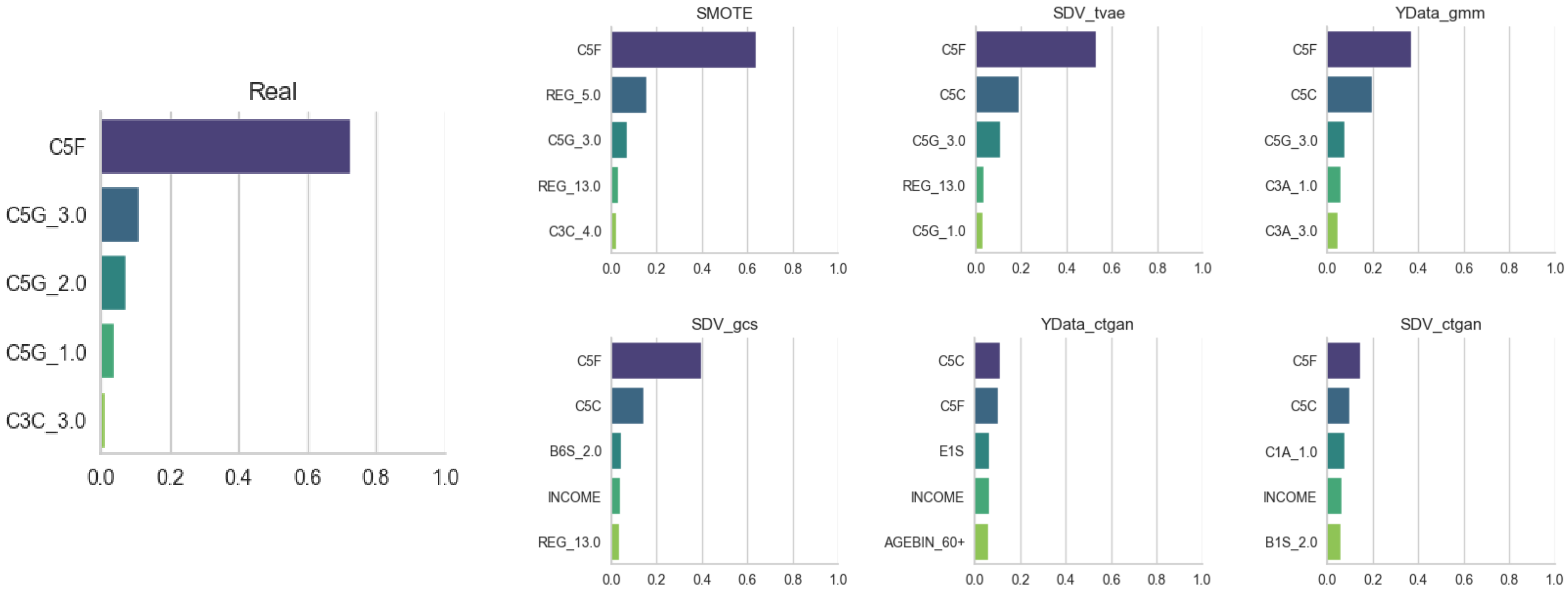
A multi-class classifier is built to **predict the range of inflation rate in the next 12 months.** Results presented are in terms of percentage difference against the real dataset.

Data		Acc.	AUC	Recall	Prec.	F1	Kappa	MCC	Average Difference
Real		***0.7852***	***0.9493***	***0.7852***	***0.7913***	***0.786***	***0.7303***	***0.7313***	
SMOTE		-0.04	-0.02	-0.04	-0.04	-0.04	-0.05	-0.05	-0.04
SDV	TVAE	-0.06	-0.03	-0.06	-0.07	-0.06	-0.08	-0.08	-0.06
YData	GMM	-0.07	-0.05	-0.07	-0.08	-0.07	-0.09	-0.09	-0.07
SDV	GC	-0.36	-0.21	-0.36	-0.41	-0.40	-0.48	-0.47	-0.38
SDV	CTGAN	-0.35	-0.24	-0.35	-0.37	-0.40	-0.46	-0.45	-0.37
YData	CTGAN	-0.52	-0.41	-0.52	-0.56	-0.56	-0.72	-0.72	-0.57

Data Utility

MACHINE LEARNING PERFORMANCE - FEATURE IMPORTANCE

The **top predictors** of inflation rate in the next 12 months are shown using feature importance scores.



Classification: GENERAL

Data Privacy

MEMBERSHIP INFERENCE

A binary classifier is built to **distinguish real (0) from synthetic (1) data** and **evaluate using precision metric or privacy score.** A low score implies an increased risk of inference, compromising individual record privacy, while a high score suggests that an attacker is unlikely to determine if a record was part of the real dataset.

Python Library	Algorithm	Accuracy	AUC Score	Precision (Privacy Score)
YData	CTGAN	0.9660	0.9972	0.9754
SDV	CTGAN	0.9451	0.9929	0.9574
SDV	GC	0.9148	0.9796	0.9260
SDV	TVAE	0.8181	0.9246	0.8472
YData	GMM	0.7911	0.9032	0.8199
In-house	SMOTE	0.7664	0.7784	0.7134

Best Synthetic Dataset

To determine the best synthetic dataset, each algorithm shall be evaluated according to this metric.

	Data Fidelity	Data Utility	Data Privacy	Overall
3 - Excellent	Has a Statistical Similarity score of 0.95 and up	Overall average difference of 0.05 or less No difference higher than 0.05 in any metric	Privacy score is higher than 0.9	Synthetic data can be used for research
2 - Fair	Statistical Similarity score is higher than 0.90 but lower than 0.95	Overall difference of 0.05 to 0.10 No difference higher than 0.10 in any metric	Privacy score is higher than 0.8 but lower than 0.9	Can be used for research but with conditions
1 - Poor	Has a Statistical Similarity score of 0.90 and lower	Difference of more than 0.10 Other conditions not satisfying any of the above	Privacy score is lower than 0.8	Not valid for research use Re-evaluate algorithm

Best Synthetic Dataset

The best synthetic dataset should have a score of 3 on all metrics. A data being produced by an algorithm having a score of 1 in any metric should not be used for research and should be re-evaluated.

3	Excellent	2	Fair	1	Poor

		In-house	YData		Synthetic Data Vault		
Metric		**SMOTE**	**CTGANs**	**Gaussian Mixture Model (GMM)**	**CTGANs**	**Gaussian Copula**	**TVAE**
Data Fidelity	Statistical Similarity	2	1	3	1	2	3
Data Utility	Machine Learning	3	1	2	1	1	2
Data Privacy	Membership Inference	1	3	2	3	3	2
AVERAGE SCORE		**2.0**	**1.7**	**2.3**	**1.7**	**2.0**	**2.3**

Key Findings and Future Works

Key Takeaways:

- Synthetic data could **replicate real data.** This can serve as an alternative and be shared with external parties. A rubric is created to decide if a synthetic data can be used for research purposes.
- For the CES dataset, synthetic datasets generated using the **TVAE and GMM** algorithm produced the best results. On the other hand, GAN-based models performed poorly in all synthetic evaluation metrics except data privacy.
- By **utilizing open-source libraries**, the implementation of generating synthetic data is much easier.

Future Works:

- Expand this study by adding numerical and time-series datasets
- Explore more algorithms for synthetic data generation
- Operationalize the synthetic data generation pipeline for research use

Thank you!

Chelsea Anne S. Ong
ongcs@bsp.gov.ph
Department of Economic Statistics
Bangko Sentral ng Pilipinas