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

CHELSEA ANNE ONG
CARMELITA ESCLANDA-LO
GABRIEL MASANGKAY
ROSSVERN REYES

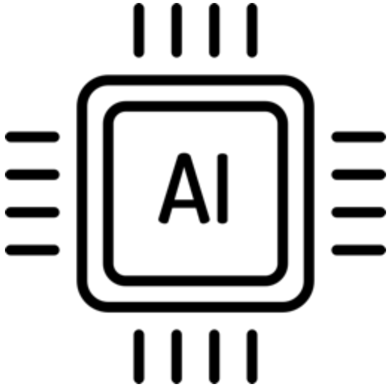
3RD IFC AND BANK OF ITALY WORKSHOP ON "DATA SCIENCE IN CENTRAL BANKING: ENHANCING THE ACCESS TO AND SHARING OF DATA", OCTOBER 17-19, 2023

OUTLINE

01 MOTIVATION
02 RELATED STUDIES
03 DATA
04 METHODOLOGY
05 RESULTS
06 KEY TAKEWAYS AND FUTURE WORK

Motivation

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."

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

01

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.

02

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)**

03

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.**

Classification: GENERAL

Synthetic Data Evaluation

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

Two types of Synthetic data:

1. **Fully Synthetic**
2. **Partially Synthetic – only identifier variables are processed**

Data Fidelity

- Statistical Similarity
  - Histogram
  - Correlation

Data Utility

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

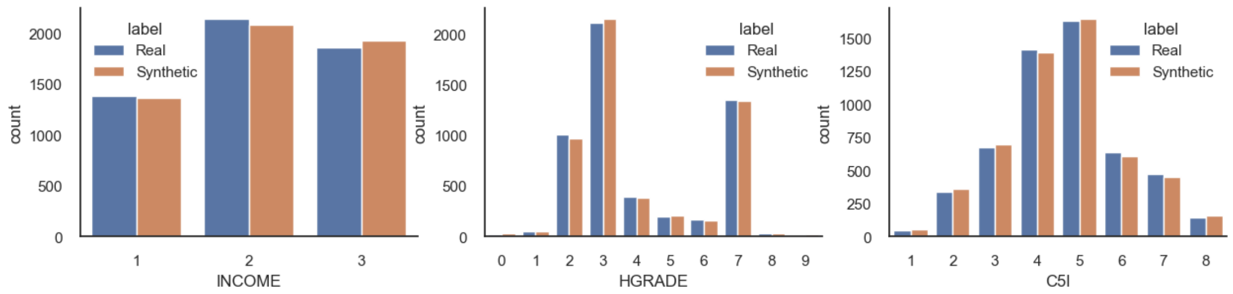
Data Privacy

- Membership Inference
  - Accuracy
  - AUC
  - Precision

Data Fidelity

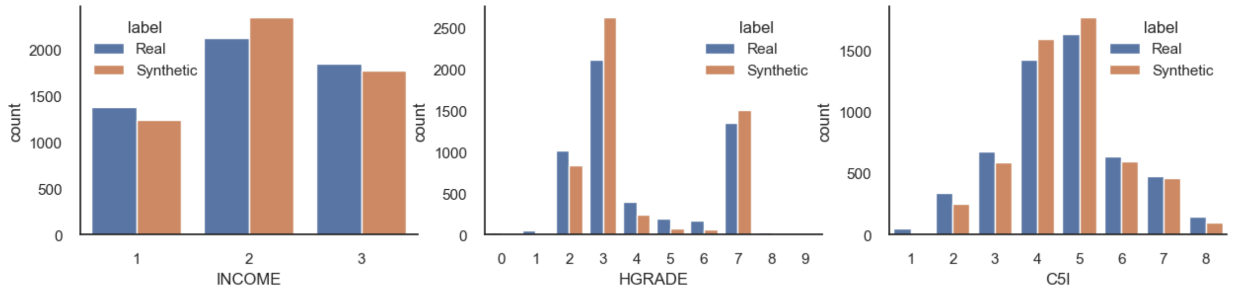
HISTOGRAM COUNTPLOTS

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)

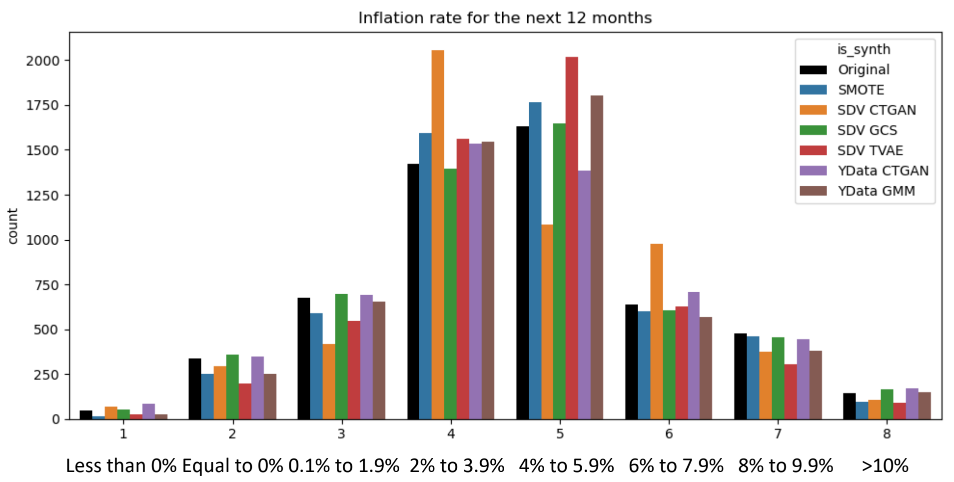


SMOTE

Similarity score = 0.8882 (± 0.0599)

**Values in parentheses are standard deviations*

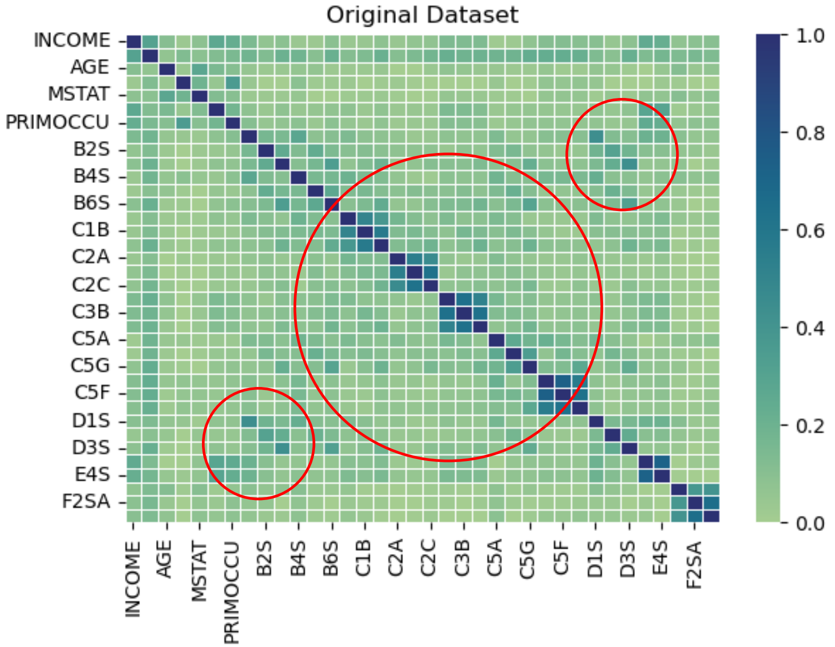
TARGET VARIABLE (C5I)

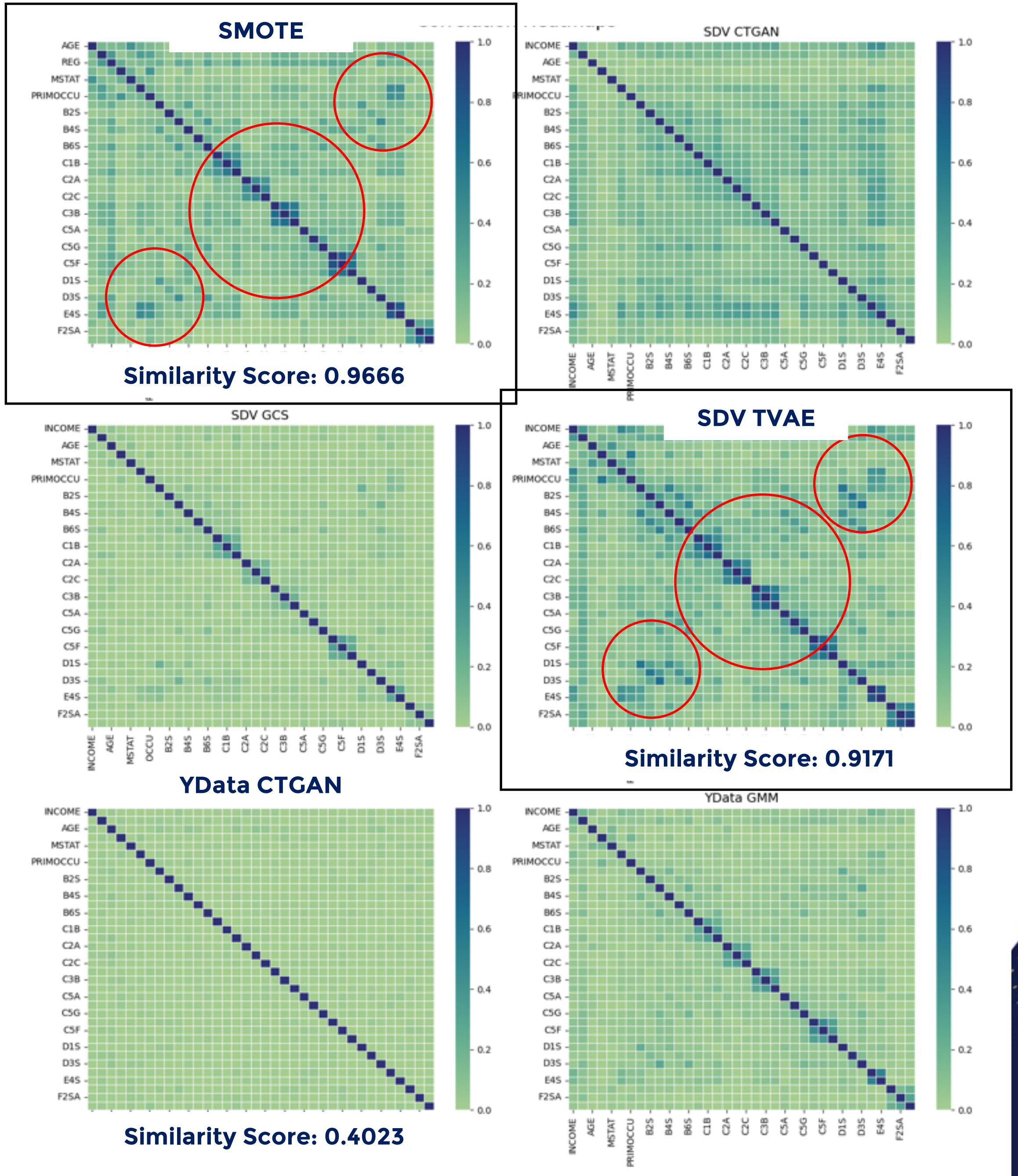


Data Fidelity

CORRELATION HEATMAP – CRAMER'S V

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





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

FULL

In-house	SMOTE	0.9896 (± 0.0259)	0.9963	0.993
SDV	GC	0.9969 (± 0.0079)	0.9091	0.953
YData	GMM	0.9888 (± 0.0299)	0.9134	0.9511
SDV	TVAE	0.9827 (± 0.0447)	0.9183	0.9505
YData	CTGANs	0.9828 (± 0.0422)	0.9133	0.9481
SDV	CTGANs	0.9795 (± 0.0451)	0.9098	0.9447

PARTIAL

**Values in parentheses are standard deviations*

Data Utility

MACHINE LEARNING PERFORMANCE

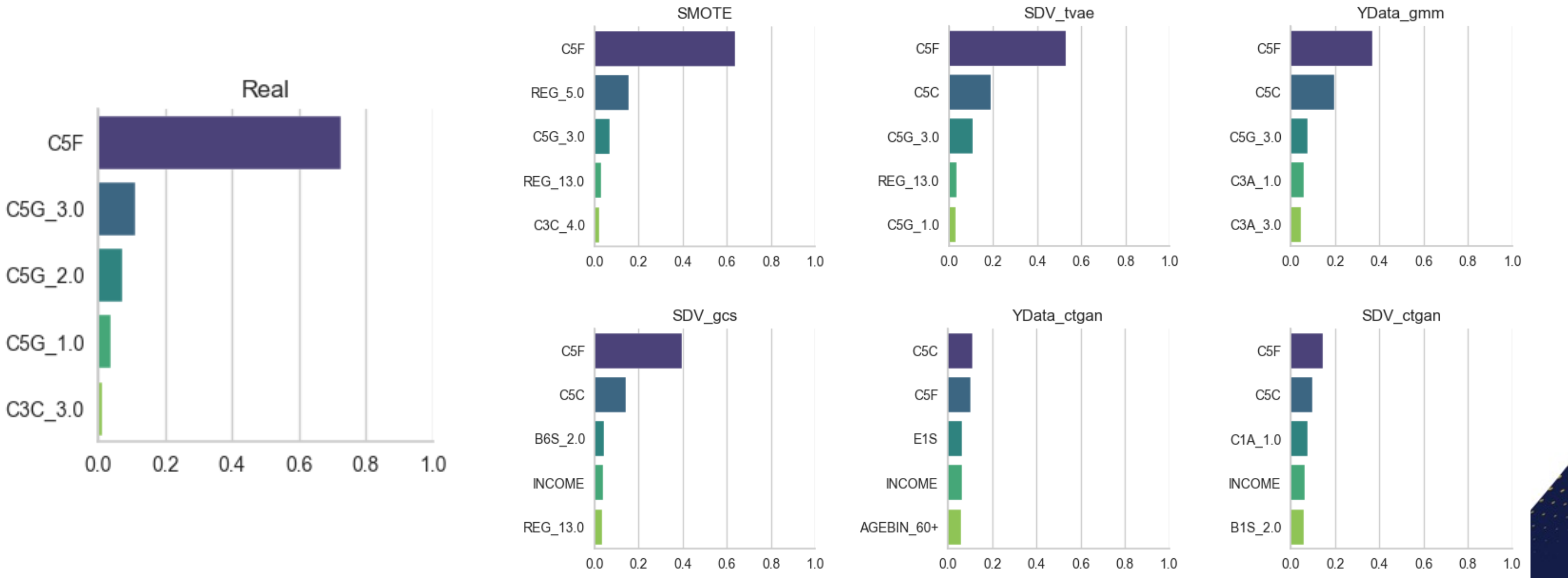
A multi-class classifier is built to **predict the range of inflation rate in the next 12 months.** Results presented are in terms of percentage difference against the real dataset.

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***	
FULL									
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
PARTIAL									
SMOTE		-0.0009	0.0006	-0.0009	-0.0011	-0.0010	-0.0011	-0.0009	-0.0008
SDV	CTGAN	-0.0101	-0.0051	-0.0101	-0.0099	-0.0100	-0.0123	-0.0122	-0.0100
SDV	GC	-0.0160	-0.0057	-0.0160	-0.0124	-0.0158	-0.0200	-0.0191	-0.0150
YData	CTGAN	-0.0196	-0.0073	-0.0196	-0.0172	-0.0196	-0.0241	-0.0234	-0.0187
SDV	TVAE	-0.0181	-0.0077	-0.0181	-0.0162	-0.0184	-0.0222	-0.0217	-0.0175
YData	GMM	-0.0229	-0.0094	-0.0229	-0.0257	-0.0234	-0.0292	-0.0295	-0.0233

Data Utility

MACHINE LEARNING PERFORMANCE - FEATURE IMPORTANCE

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



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.

FULL

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

PARTIAL

Python Library	Algorithm	Accuracy	AUC Score	Precision (Privacy Score)
YData	CTGAN	0.8027	0.8858	0.8022
SDV	CTGAN	0.7548	0.8437	0.7591
SDV	GC	0.7617	0.8465	0.7587
SDV	TVAE	0.7520	0.8232	0.7406
YData	GMM	0.7329	0.8057	0.7215
In-house	SMOTE	0.4984	0.1987	0.2577

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

Metric		In-house	YData		Synthetic Data Vault		
		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.
- **Partially synthetic data sets** can be used for research purposes but with conditions (i.e., only share data internally).
- 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