Stonk Bench: Unified Benchmark for Synthetic Data Generation for Financial Time Series

Uyen Lam Ho*
Eddison Pham*
uyenlam.ho@mail.utoronto.ca
eddison.pham@mail.utoronto.ca
University of Toronto
Toronto, Ontario, Canda

Figure 1: Toronto Stock Exchange, Toronto, 1981

Abstract

The ever emerging field of Synthetic Data Generation (SDG) has gain traction within many financial domains, including Finacial Time Series (FTS) data [3]. However, there has been discorse and lack in universal agreement in what determines a better SDGFTS (Synthetic Data Generation for Financial Time Series), which maybe hindering progress in the field [4]. In this paper, we propose and contribute a comprehensive benchmark for SDGFTS, covering various aspects such as data quality, privacy, and utility. We evaluate several state-of-the-art SDG methods using our benchmark and provide insights into their strengths and weaknesses. Our first-of-its-kind benchmark aims to facilitate the development of more effective SDG methods for FTS data, incorperating both classical statistical measures and utility benefits, and test over (insert number) to decicively determine what is the best SDGFTS model right now and for the future.

Keywords

Synthetic Data Generation, Financial Time Series, Benchmarking

ACM Reference Format:
Uyen Lam Ho and Eddison Pham. 2025. Stonk Bench: Unified Benchmark for Synthetic Data Generation for Financial Time Series. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX).* ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Synthetic Data Generation (SDG) has emerged as a crucial tool in financial technology, particularly for Financial Time Series (FTS) data [3]. The ability to generate high-quality synthetic financial data addresses several critical challenges in the field, including data privacy concerns, limited data availability, and the need for diverse training datasets in machine learning applications [4].

Despite the growing importance of SDGFTS (Synthetic Data Generation for Financial Time Series), there is a notable lack of standardization in evaluating the quality and effectiveness of these generative models [4]. Current evaluation methods vary widely across studies, making it difficult to compare different approaches objectively and determine their relative strengths and weaknesses.

Our research addresses this gap by introducing a comprehensive benchmark framework that encompasses:

- Statistical fidelity measures for comparing synthetic and real FTS
- Privacy preservation metrics to ensure sensitive financial information remains protected

[*] Both authors contributed equally to this research.

Conference acronym 'XX, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

- Utility metrics that assess the practical value of synthetic data in downstream tasks
- Performance evaluation across multiple SDGFTS models and datasets

By analyzing the results from our MLflow experiments and applying our unified benchmark, we aim to provide clear guidelines for evaluating SDGFTS models and establish a standard for future research in this domain.

1.1 Limitations in the SDGFTS Literature

Among other things, we observed that there is currently no universal, generally accepted approach to evaluating synthetic time series[4]; this issue extends beyond FTS to generative frameworks like GANs as a whole [5]. Many evaluation measures are insufficiently defined and lack public implementations, making reuse and reproduction troublesome and error-prone [4]. This presents a challenge unique to the generation task compared to areas like time series forecasting or classification [4]. Hence, future research would immensely benefit from a widely accepted, reasonably sized set of qualified measures for the central evaluation criteria.

1.2 Our Contributions

To put it in simple terms, we are adopting the time series evaluation taxonomy outlined from Stenger et al. [4], develop a comprehensive and unified SDG benchmark expanded from the works of Ang et al. [1] in specifics to FTS by incorporating a utility evaluation framework inspired by Boursin et al. [2] through portfolio evaluations generated by a deep hedger. Our key contributions include:

[C1] **Consolidated Evaluation Framework:** We systematically review and consolidate evaluation methods from leading papers (models and surveys) in SDG and financial time series [1, 2, 4], creating a comprehensive assessment toolkit based on established practices.

[C2] **Unified Statistical and Utility Measures:** For the first time, we integrate both statistical fidelity metrics and practical utility measures in a single SDGFTS benchmark, providing a more complete evaluation of synthetic data quality.

[C3] **Open Benchmark Platform:** We deliver an open-source benchmark framework for the research community, aiming to establish a gold standard for SDGFTS model evaluation and comparison.

2 Preliminaries

2.1 Problem Definition

Let us formally define the Synthetic Data Generation (SDG) problem for Financial Time Series (FTS). Given a financial time series X with N individual series of length T, we represent it as a matrix:

$X = (x_1, ..., x_n)^T$

where each individual series x_i is a T-dimensional vector:

$x_i = (x_{i,1}, ..., x_{i,t})$

and each $x_{i,t}$ corresponds to a single time point t of x_i.

Let $P(x_1, ..., x_n)$ denote the real distribution of the given time series X. The objective of SDG for FTS is to generate a synthetic time series:

$\hat{X} = (\hat{x}_1, ..., \hat{x}_n)$

such that its distribution $P(\hat{x}_1, ..., \hat{x}_n)$ approximates $P(x_1, ..., x_n)$, while preserving key statistical properties and financial characteristics of the original data. These properties include:

2.2 Scope of Project

2.2.1 *Scope of Methods.*

Our benchmark encompasses a diverse range of SDGFTS methods, from traditional parametric approaches to modern deep learning architectures. We selected models based on three main criteria: (1) proven industry adoption, (2) research community acceptance, and (3) recent methodological innovations.

Our evaluation includes classical parametric models, which rely on predefined statistical distributions and assumptions about the underlying data generation process. These models have been extensively used in financial institutions for their interpretability and theoretical foundations.

We also evaluate modern non-parametric approaches, particularly deep learning-based models that learn the data distribution directly from observations without assuming a specific form. These include generative adversarial networks, variational autoencoders, and diffusion models, representing the cutting edge in synthetic data generation.

2.2.2 *Scope of Datasets.*

We evaluate models on diverse financial datasets spanning:

- Stock market indices (S&P 500, NASDAQ, TSX)
- Individual stock prices from major exchanges
- Forex trading pairs

2.2.3 *Scope of Evaluation Taxonomical Criteria.*

Our evaluation framework follows the taxonomy structure proposed by Stenger et al. [4], incorporating various evaluation measures from different papers in the field.

We systematically integrate evaluation metrics from multiple sources while maintaining this structured taxonomical approach, ensuring comprehensive coverage of all critical aspects of SDGFTS evaluation and setting a standard for future research in time series SDG as a whole.

3 Overview of FTS methods

3.1 Classical Statistical (Parametric) Methods

3.2 Deep Learning-based (Non-parametric) Methods

3.2.1 *Generative Adversarial Networks (GANs).*

3.2.2 *Variational Autoencoders (VAEs).*

3.2.3 *Diffusion Models.*

3.2.4 *Others.*

4 Stonk Bench Architechture

4.1 Datasets and Preprocessing

4.2 Statistical Evaluation Measures

4.2.1 *Feature-based Distance Measures.*

These metrics quantify how well synthetic data captures the statistical properties of real data:

[M1] **Marginal Distribution Difference (MDD):** Measures the overall difference between the probability distributions of real and synthetic data, helping assess if the synthetic data maintains the same value ranges and frequencies.

[M2] **Mean Difference (MD):** Compares the central tendencies of real and synthetic data, ensuring the synthetic data preserves the average behavior of the time series.

[M3] **Standard Deviation Difference (SDD):** Evaluates how well the synthetic data maintains the volatility characteristics of the real data by comparing their spread measures.

[M4] **Kurtosis Difference (KD):** Assesses preservation of the "tailedness" of distributions, crucial for capturing extreme events in financial data.

[M5] **AutoCorrelation Difference (ACD):** Measures how well temporal dependencies are preserved in the synthetic data compared to real data.

4.2.2 Visualization Methods. Visual analysis tools provide intuitive validation of synthetic data quality and visual interpretive comparisons and contrast between real and synthetic time series [1]. Visual assessment is also useful for presentation purposes when presenting to non-technical stakeholders.

[M6] **t-SNE Visualization:** Reduces high-dimensional financial data to 2D representations, allowing visual comparison of real and synthetic data structure.

[M7] **Distribution Comparison Plots:** Direct visual comparison of probability distributions between real and synthetic data through histograms and Q-Q plots.

4.2.3 Diversity Metrics. These metrics ensure synthetic data provides meaningful variations:

[M8] **Euclidean Distance (ED):** Measures basic similarity between synthetic samples, helping assess diversity within generated data.

[M9] **Dynamic Time Warping (DTW):** Captures temporal similarities while allowing for slight shifts and warps in time series patterns.

4.2.4 Efficiency Assessment.

[M10] **Generation Time:** Measures computational efficiency by tracking time required to generate 500 synthetic samples, crucial for practical applications.

4.2.5 Stylized Facts Verification. These tests verify if synthetic data exhibits key properties of financial time series:

[M11] **Heavy Tails:** Measures excess kurtosis to verify if synthetic data captures the frequent extreme events characteristic of financial returns.

[M12] **Return Autocorrelation:** Checks for proper modeling of temporal dependencies in returns through lag-1 autocorrelation.

[M13] **Volatility Clustering:** Verifies if periods of high volatility tend to cluster together, a key feature of financial markets.

[M14] **Long Memory in Volatitly:** Tests for persistent autocorrelation in absolute returns, indicating proper modeling of volatility persistence.

[M15] **Non-Stationarity Detection:** Ensures synthetic data maintains the non-stationary characteristics typical of financial time series.

4.3 Utility Evaluation Measures: Deep Hedging

Utility measures are critical for evaluating synthetic data beyond statistical metrics, as they assess the practical value of generated data in real-world financial applications [2]. Our benchmark incorporates deep hedging as a utility measure for several key reasons:

[U1] **Real-world Application Testing:** Deep hedging provides a concrete way to evaluate how synthetic data performs in actual financial tasks, particularly in derivatives pricing and risk management.

[U2] **Industry-relevant Metrics:** By comparing hedging strategies trained on synthetic versus real data, we can assess the practical utility of synthetic data through metrics that matter to financial practitioners, such as replication errors and hedging performance.

[U3] **Model Robustness Validation:** Deep hedging helps verify if synthetic data maintains the complex relationships and market dynamics necessary for developing reliable trading strategies.

4.3.1 Deep Hedger Problem Defintion.

4.3.2 Deep Hedger Architecture.

4.3.3 Deep Hedger Portfolio Evaluation.

5 Results and Analysis

5.1 Statistical Evaluation Results

5.2 Utility Evaluation Results

5.3 Comprehensive Model Comparison

5.4 Ranking Analysis

6 Conclusion and Future Work

In our research, we recognize the existing limitations in the evaluation of synthetic time series, particularly the absence of a universally accepted framework. To address this, we adopt the evaluation taxonomy proposed by Stenger et al. and expand upon it to create a comprehensive benchmark specifically tailored for Synthetic Data Generation for Financial Time Series (SDGFTS). Our contributions include the development of a consolidated evaluation framework that systematically reviews and integrates various assessment methods from leading studies in the field. This approach not only enhances the rigor of evaluations but also facilitates the comparison of different SDGFTS models, ultimately providing a more standardized and reliable means of assessing synthetic data quality.

6.1 Summary of Findings

6.2 Implications for SDGFTS Research

6.3 Current Limitations and Shortcomings

6.4 Future Research Directions

Acknowledgments

To professor Irene Huang, University of Toronto, for her supervision and guidance throughout the project.

References

[1] Yihao Ang, Qiang Huang, Yifan Bao, Anthony K. H. Tung, and Zhiyong Huang. 2023. TSGBench: Time Series Generation Benchmark. *Proc. VLDB Endow.* 17, 3 (2023), 305–318.

[2] Nicolas Boursin, Carl Remlinger, Joseph Mikael, and Carol Anne Hargreaves. 2022. Deep Generators on Commodity Markets; application to Deep Hedging. arXiv:2205.13942 [q-fin.RM] https://arxiv.org/abs/2205.13942

[3] Vamsi K. Potluru, Daniel Borrajo, Andrea Coletta, Niccolo Dalmasso, Yousef El-Laham, Elizabeth Fons, Mohsen Ghassemi, Sriram Gopalakrishnan, Vikesh Gosai, Eleonora Kreacic, Ganapathy Mani, Saheed Obitayo, Deepak Paramanand, Natraj Raman, Mikhail Solonin, Srijan Sood, Svitlana Vyetrenko, Haibei Zhu, Manuela Veloso, and Tucker Balch. 2024. Synthetic Data Applications in Finance. arXiv:2401.00081 [cs.LG] https://arxiv.org/abs/2401.00081

[4] Michael Stenger, Robert Leppich, Ian Foster, Samuel Kounev, and André Bauer. 2024. Evaluation is key: a survey on evaluation measures for synthetic time series. *J Big Data* 11, 66 (May 2024). doi:10.1186/s40537-024-00924-7

[5] Zhengwei Wang, Qi She, and Tomas E. Ward. 2020. Generative Adversarial Networks in Computer Vision: A Survey and Taxonomy. arXiv:1906.01529 [cs.LG] https://arxiv.org/abs/1906.01529