SAD Money: Using survival analysis divergence to evaluate and refine financial time series generation

Keywords: Synthetic Data, Financial Time Series, Model Evaluation

1 Introduction

Synthetic financial time series generation has emerged as a promising tool for enhancing portfolio management strategies. While much of the current focus in this field is on applications for high-frequency trading firms, such as hedge funds, these methods are less explored in the context of longer-term portfolio management. This gap is critical because the majority of individual investors prioritize strategies spanning months or years, rather than milliseconds. Generative adversarial networks (GANs) have shown potential in generating realistic financial data, but their use is often tailored to short-term trading and lacks rigorous evaluation metrics for longer-term trends (Dogariu et al., 2022; Liao et al., 2024). Current evaluation methods often emphasize statistical similarities, such as matching marginal distributions or pairwise correlations, but fail to address practical, investor-relevant dynamics, such as event fidelity (e.g., accurately reproducing major market shifts) and the timing of critical price thresholds (Theis et al., 2015). Bridging this gap requires rethinking model training approaches and developing evaluation techniques tailored to the needs of long-term investors.

We propose a framework for financial time series generation centered around a novel evaluation metric to assess synthetic data. Our approach is a start at bridging the gap between current data generation methods and individual investors' needs, and aims to enable stochastic backtesting of long-term portfolio strategies via higher-fidelity synthetic data.

2 Methods

Let τ denote the time period of interest and S the set of financial instruments. The observed time series, $X_{\tau,S}$, represents a single realization from the broader population $\mathcal{X}_{\tau,S}$, encompassing all possible price trajectories over τ. The goal of a data synthesizer is to generate $\tilde{X}_{\tau,S} \in \mathcal{X}_{\tau,S}$ while preserving the statistical properties and dynamics of the population.

2.1 Mapping to Survival Distributions

To evaluate the quality of the synthetic data $\tilde{X}_{\tau,S}$, we map both real and synthetic data to survival distributions. Let $\alpha \in \mathcal{A} \subseteq [-1, \infty)$ represent a chosen cumulative percentage change (e.g., -5% or 18%), with $\mathcal{A}$ selected based on the financial instruments and time period. For each instrument in S, we compute cumulative percentage changes over τ and define the event of interest as the first time α is reached. This process yields event times for α, allowing construction of survival distributions, with right censoring for instruments that do not reach α during τ.

The survival distribution $\bar{F}_{\alpha,S,\tau}(t)$ gives the probability of not hitting α by time t, while $\bar{F}^*_{\alpha,S,\tau}(t)$ represents the same for the synthetic data.

2.2 Survival Analysis Divergence

The Survival Analysis Divergence (SAD) quantifies the discrepancy between the survival distributions of the real and synthetic data. This metric provides a holistic summary of differences across both time and cumulative percentage changes. Formally, the SAD is defined as

$$\text{SAD} = \int_{\alpha \in \mathcal{A}} \int_{t \in \tau} w(\alpha, t) |\bar{F}^*_{\alpha,S,\tau}(t) - \bar{F}_{\alpha,S,\tau}(t)| \, dt \, d\alpha,$$

where $w(\alpha, t)$ is taken to be a weighting function dependent on both α and t.

The SAD integrates over all cumulative percentage changes $\alpha \in \mathcal{A}$ and all time points $t \in \tau$, capturing the total deviation of the synthetic survival distribution from the real distribution. Taking $w(\alpha, t)$ to be a constant means that discrepancies are evaluated uniformly across a wide range of events and time scales, while varying its values based on t and α is akin to giving higher weights to certain discrepancies.

3 Results

We propose a framework for generating synthetic financial time series that maintains the overall trajectory of the market, as well as intra-stock and inter-stock correlations. Currently, to the best of our knowledge, no publicly available models exist that can simultaneously generate such data. For our analysis, we utilized synthetic data from two different sources: the J.P. Morgan AI Research Synthetic Dataset: Equity Markets Data[1] (Wiese et al., 2019; Liao et al., 2024) and the Skanalytix Synthetic Financial Time Series Generator (Skabar, 2024).

The J.P. Morgan dataset is not tied to any specific time period, making it necessary to introduce a trended approach to align the synthetic data with historical trends. To achieve this, we added the average daily returns of the stocks in S to the synthetic returns. Additionally, the dataset does not claim to preserve inter-stock correlations, and the generated trajectories were not informed by the real sample of stocks used in this analysis.

In contrast, the Skanalytix Synthetic Financial Time Series Generator samples from input distributions and generates time series informed by the returns of each asset in S over the specified time period τ. While some inter-stock correlations were preserved (as up to four series could be generated simultaneously), the timing of broader market fluctuations was not captured in the synthetic data. Consequently, we analyzed this dataset in both its raw state and a trended state, similar to the J.P. Morgan data.

For our experiments, we used a random sample of 52 stocks from the S&P 500 index and considered three distinct time periods, τ: 2015, 2020, and 2022. The range of percentage changes analyzed was $\mathcal{A} = [-0.3, 0.3]$, and computations

[1] This publication includes or references synthetic data provided by J.P. Morgan.

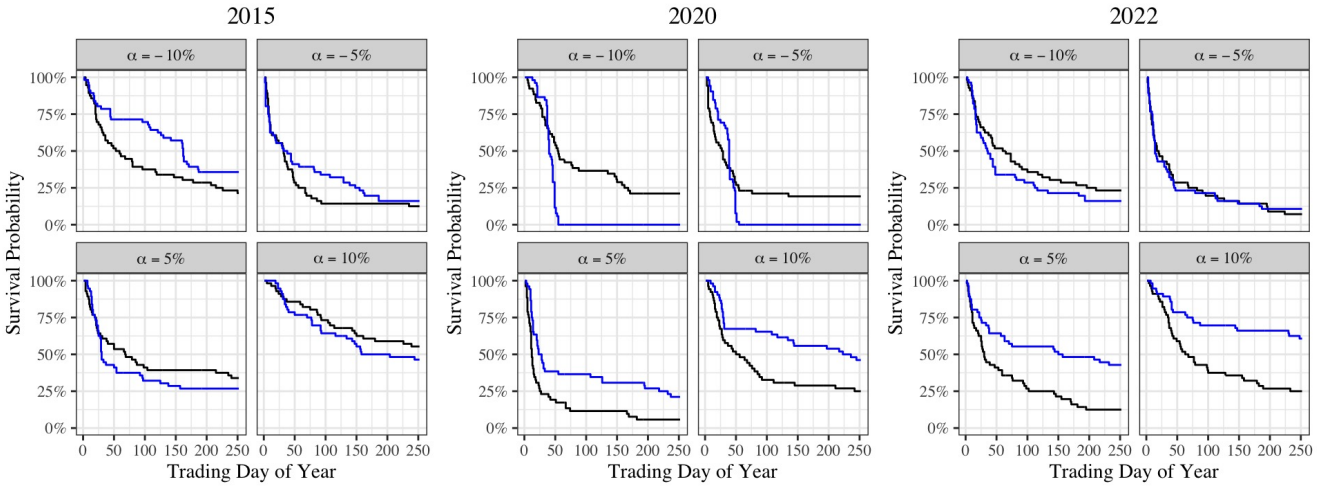

Figure 1: Examples of survival distributions for different α values across the years of interest for the Skanalytix Raw dataset. The survival function for the observed data can be seen in blue, while the survival function for the synthetic data is in black.

were carried out numerically on a grid with a spacing of 0.01. When computing the Survival Analysis Divergence (SAD) for these periods, we employed a constant weighting function of $w(\alpha, t) = (\|\tau\| \cdot \|\mathcal{A}\|)^{-1}$, where $\|\cdot\|$ represents the cardinality of the set. This allowed the SAD in our application to be interpreted as the average absolute difference in survival probability across τ and $\mathcal{A}$.

The computed SAD values across the different years and datasets are summarized in Table 1. Additionally, selected examples of survival distributions are shown in Figure 1.

Dataset	Reference Year		
	2015	2020	2022
J.P. Morgan Raw	0.0460	0.2177	0.1227
J.P. Morgan Trended	0.0609	0.0657	0.0814
Skanalytix Raw	0.1013	0.2638	0.1394
Skanalytix Trended	0.1200	0.1228	0.1041

Table 1: Survival Analysis Divergence (SAD) values for various synthetic datasets compared to real data across selected years. Smaller values indicate more accurate synthetic data.

4 Discussion

The validity of synthetic data, as measured by the SAD, varies significantly across datasets and years. Trended approaches consistently outperform raw approaches in years with clear time-dependent trends, such as 2020, where the raw synthetic data fails to capture the significant market downturn caused by the COVID-19 pandemic (Figure 1). In contrast, trended data better reflects this critical event, highlighting the importance of incorporating trends into synthetic data generation. The J.P. Morgan synthetic data performs the best in 2015, likely because its zero-growth centering aligns with the relatively flat market conditions of that year. This similarity between the synthetic data's characteristics and the actual market behavior results in a lower SAD.

On average, we propose that an absolute difference in survival probability of 5% is a reasonable benchmark for synthetic data quality. However, among the synthetic datasets available to us, this goal is achieved only once. If synthetic data is to become a reliable tool for improving portfolio management strategies, further effort is needed to reduce the observed discrepancies between synthetic and real data.

Several directions for future research aim to enhance the generation of synthetic financial time series. A primary goal is to explore more complex and meaningful weighting functions, $w(\alpha, t)$, that better reflect real-world financial considerations. Additionally, the SAD metric itself could serve as a training objective for generative adversarial networks (GANs). Specifically, incorporating the SAD as part of the discriminator's task in a GAN framework may lead to the generation of more realistic synthetic data. By using the SAD to directly inform the training process, this approach has the potential to significantly improve the quality of generated samples.

References

Mihai Dogariu, Liviu-Daniel Ştefan, Bogdan Andrei Boteanu, Claudiu Lamba, Bomi Kim, and Bogdan Ionescu. 2022. Generation of Realistic Synthetic Financial Time-series. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 18(4):1–27, November.

Shujian Liao, Hao Ni, Marc Sabate-Vidales, Lukasz Szpruch, Magnus Wiese, and Baoren Xiao. 2024. Sig-Wasserstein GANs for conditional time series generation. *Mathematical Finance*, 34(2):622–670, April.

Andrew Skabar. 2024. Generating Realistic Synthetic Financial Time Series, August.

Lucas Theis, Aäron van den Oord, and Matthias Bethge. 2015. A note on the evaluation of generative models. *arXiv preprint arXiv:1511.01844*.

Magnus Wiese, Lianjun Bai, Ben Wood, and Hans Buehler. 2019. Deep hedging: learning to simulate equity option markets. *arXiv preprint arXiv:1911.01700*.