Quantitative Finance

ISSN: 1469-7688 (Print) 1469-7696 (Online) Journal homepage: www.tandfonline.com/journals/rquf20

Generation of synthetic financial time series by diffusion models

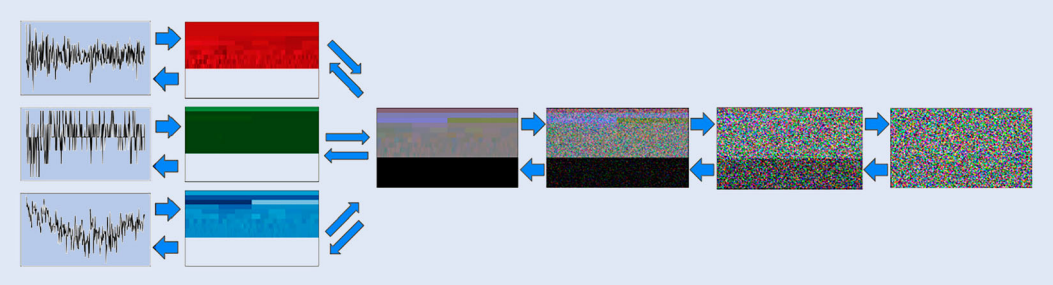
Tomonori Takahashi & Takayuki Mizuno

To cite this article: Tomonori Takahashi & Takayuki Mizuno (05 Aug 2025): Generation of synthetic financial time series by diffusion models, Quantitative Finance, DOI: 10.1080/14697688.2025.2528697

To link to this article: https://doi.org/10.1080/14697688.2025.2528697

Published online: 05 Aug 2025.

Quantitative Finance, 2025
https://doi.org/10.1080/14697688.2025.2528697



Generation of synthetic financial time series by diffusion models

TOMONORI TAKAHASHI †† and TAKAYUKI MIZUNO ‡‡

†The Graduate University for Advanced Studies, SOKENDAI, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan
‡National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan

(*Received 21 October 2024; accepted 26 June 2025*)

Despite its practical significance, generating realistic synthetic financial time series is challenging due to statistical properties known as stylized facts, such as fat tails, volatility clustering, and seasonality patterns. Various generative models, including generative adversarial networks (GANs) and variational autoencoders (VAEs), have been employed to address this challenge, although no model yet satisfies all the stylized facts. We alternatively propose utilizing diffusion models, specifically denoising diffusion probabilistic models (DDPMs), to generate synthetic financial time series. This approach employs wavelet transformation to convert multiple time series (into images), such as stock prices, trading volumes, and spreads. Given these converted images, the model gains the ability to generate images that can be transformed back into realistic time series by inverse wavelet transformation. We demonstrate that our proposed approach satisfies stylized facts.

Keywords: Finance; Time series; Synthetic data; Diffusion models; Econophysics

JEL Classification: C22, C32, C45, C63, G17

1. Introduction

Financial markets generate a wide range of time series data every day, including stock prices, trading volumes, and bid-ask spreads. These series are important both for practitioners, who use them for risk management and trading strategies, and for academics not only in economics but also in statistics and physics, who seek to model and understand complex market behavior, which consists of a large number of interacting agents (Plerou *et al.* 2000, Chakraborti *et al.* 2007). A common starting point for studying stock prices is to model them as random walks or Brownian motions (Bachelier 1900), in which case, taking log returns (or differences) yields a stationary time series.

However, a considerable body of research has demonstrated that real financial time series deviate from this simple framework, exhibiting so-called 'stylized facts.' These include heavy (fat) tails in the return distributions (Gabaix *et al.* 2003, Gabaix 2009), volatility clustering (Cont 2001, Chakraborti *et al.* 2007, Ratliff-Crain *et al.* 2023), and seasonal and intraday patterns (Shakeel and Srivastava 2018). Empirical evidence suggests that the return distribution often exhibits a power-law decay, while volatility and other market variables (e.g. spreads, trading volumes) frequently show long memory. Although parametric methods (e.g.ARCH-type models) (Bollerslev *et al.* 1992, Takayasu *et al.* 2010) and agent-based models (Lux 2009, Samanidou *et al.* 2007) have been developed to reproduce or explain these phenomena, they often focus on uncovering the mechanisms or sources of the stylized facts rather than generating realistic time series in full detail.

In parallel, an important research direction is the generation of synthetic financial time series that capture the statistical properties observed in real markets. Such synthetic data can be highly useful for risk modeling and stress-testing, particularly in rare market conditions (Eckerli and Osterrieder 2021, Brophy *et al.* 2023). Recently, machine learning-based generative models such as Generative Adversarial Networks (GANs) (Goodfellow *et al.* 2014) and Variational Autoencoders (VAEs) (Kingma and Welling 2014) have shown promise in generating realistic samples that capture many stylized facts (Wiese *et al.* 2020, Eckerli and Osterrieder

*Corresponding author. Email: t_takahashi@nii.ac.jp

2021, Dogariu *et al.* 2022). However, fully reproducing all major stylized facts (fat tails, volatility clustering, seasonality, and cross-correlations) remains a challenge (Dogariu *et al.* 2022).

To address these issues, we propose an alternative methodology based on Denoising Diffusion Probabilistic Models (DDPMs) (Ho *et al.* 2020), which have recently shown state-of-the-art performance in image generation particularly in terms of quality and divergence of synthetic data (Xiao *et al.* 2022). Our key insight is to represent financial time series as images using a wavelet transformation (Ramsey *et al.* 1995), converting each day's (or interval's) price returns, spreads, and trading volumes into a single color image (one channel per time series). By harnessing DDPMs' capabilities in image generation, we facilitate the simultaneous capture of both time-domain and frequency-domain information. We then invert the wavelet transformation to recover the synthetic time series. Thus, our method effectively learns and reproduces the multivariate structure of returns, spreads, and trading volumes simultaneously, capturing fat tails, volatility clustering, cross-correlations, and the well-known intraday seasonality pattern.

The remainder of the paper is organized as follows. Section 2 reviews related literature on financial time series from the perspectives of econophysics and machine learning. Section 3 presents our proposed methodology, covering the wavelet transformation, the DDPM training, and the process of reconstructing time series from generated images. Section 4 provides experimental results, demonstrating that our approach successfully reproduces multiple stylized facts. Section 5 concludes with discussions and directions for future research.

2. Related works

Time series data, a critical component for prediction problems, has received considerable attention, especially from numerous studies aimed at enhancing prediction accuracy (Das *et al.* 2023). Time series data have attracted studies in both econophysics and informatics. The former focuses on the stylized facts of financial time series; the latter informatics concentrates on generating time series data. In this section, we review related works in these fields.

2.1. Related works in econophysics

Financial time series are generated in actual financial markets every day, every minute, or even more frequently by the activities of the participants in such markets. The most typical example of a financial time series is the price movement of stocks. The spreads and trading volumes of stocks are also observed in financial markets. Here, there are two types of prices: bids and asks. The bid price refers to the buyer's interest price, and the ask price refers to the seller's interest price. If a market participant wishes to immediately purchase (resp. sell) a particular stock, she can place a request called a market order. As a result, the transaction is executed at the ask (resp. bid) price. The difference of ask and bid prices (generally, bid $<$ ask) is called a bid-ask spread or just the spread. When a transaction is executed by matching a buy market order to the ask price or a sell market order to the bid price, the amount of the transaction's stock is also recorded as time series data, and these amounts are called the trading volumes. The prices and spreads are time series created by observing their snapshots at predetermined intervals: daily, hourly, minute-by-minute, or even more frequently. The trading volumes are also time series based on summing the amounts traded in the intervals. We treat these three typical time series data as financial time series (figure 1). Regarding stock prices S_i, their log returns $\log \frac{S_i}{S_{i-1}}$ are often used instead of prices to normalize the differences among various stocks. This transformation also moves stock price fluctuations closer to a stationary state, similar to how a random walk, which is a non-stationary time series, always shows stationarity when a difference series is taken.

These financial time series, such as stock prices, resemble Brownian motions. In the early days of theoretical studies of financial time series, modeling was done using Brownian motions (Bachelier 1900). Over the past few decades, many studies have proved that financial time series exhibit unique characteristics that are not seen in Brownian motions. As explained above, the deviation of financial time series from Brownian motion (called stylized facts) has been studied extensively. One typical stylized fact is the fat-tail phenomenon, where financial time series tend to have heavier tails than a normal distribution, and they follow power laws (Gabaix *et al.* 2003, Gabaix 2009). The emergence of these larger than expected movements under a normal distribution is often attributed to the trades of large participants (Gabaix *et al.* 2003). Another stylized fact is volatility clustering, where large fluctuations tend to be followed by periods of similarly large fluctuations. This characteristic appears as the slow decay of autocorrelation of volatility time series with respect to lag. Other slow decays of autocorrelations are observed in time series of spreads and trading volumes, which interrelate with volatilities. This characteristic, called the *long memory* of time series, contrasts with the short memory of log returns of stock prices, which have the fast decay of autocorrelation. Such stylized facts are observed not only in stock markets but also in other various financial markets like foreign exchange markets (Mizuno *et al.* 2003). Studies have employed parametric models to represent stylized facts. The ARCH model is a key instrument for empirical studies (Bollerslev *et al.* 1992), and some studies have modified random walks (Takayasu *et al.* 2010). Another approach for studying stylized facts is agent-based models, which simulate market participants as agents (Samanidou *et al.* 2007, Lux 2009).

Additionally, observing the movement of financial time series throughout the day reveals intraday seasonality. Volatility, spread, and trading volume are high immediately after the market opens, decrease during mid-day, and rise again as the market approaches closing. These characteristics have been studied in the context of high-frequency market microstructure (Shakeel and Srivastava 2018).

Wavelet analysis is another approach for analyzing financial time series (Ramsey *et al.* 1995). Wavelet transformation can decompose one-dimensional time series data into two-dimensional pictures in space and time, allowing the nature of the original time series to appear in the pictures.

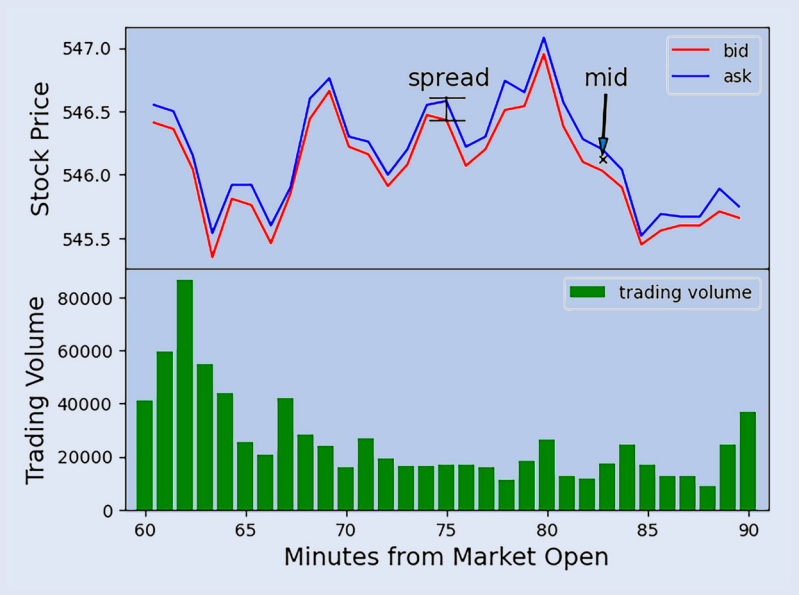


Figure 1. Examples of typical financial time series. Data: minute-based stock prices and trading volumes of AAPL.O on NASDAQ on January 23, 2014 from 10:30 to 11:30 AM EST. In a stock exchange, both bid and ask prices are observed. A bid (resp. ask) price is an amount quoted by market participants who want to buy (resp. sell) the stock. Arithmetic average price of bid and ask prices for each time are called the mid-price (the mid), and difference between ask and bid prices is called the bid-ask spread (the spread).

2.2. Related works in informatics

Attempts have explored the nature of time series and the mechanisms that cause them as well as the generation of realistic synthetic time series. For instance, in the medical and social science fields, synthetic time series can support model training through data augmentation while preserving anonymity (Brophy *et al.* 2023). In the financial sector, there is a practical interest in employing synthetic time series for stress tests under hypothetical market conditions for improving the robustness of risk management and trading models. Given the restricted access to high-frequency financial data and the scarcity of time series that demonstrate such unique phenomena as 'flash crashes,' a concerted effort is focusing on generating diverse, realistic financial time series (Wiese *et al.* 2020, Eckerli and Osterrieder 2021, Dogariu *et al.* 2022).

Generative adversarial networks (GANs) have been employed to produce not only financial time series but also other general time series. GANs were originally proposed to generate images (Goodfellow *et al.* 2014). A GAN has two types of neural network structures: a generator and a discriminator. The generator creates realistic data from random noise data. The discriminator receives real data or data from the generator and decides whether they are real or generated. If the discriminator's decision is correct, the generator incurs a loss; otherwise, the discriminator suffers a loss. By continuing this process as neural network training, the generator eventually learns to generate realistic data. TimeGAN (Yoon *et al.* 2019) is a GAN model specifically designed to generate time series data. It introduces latent space and two functions (embedding and recovery) to map the characteristics of input data into the latent space. As another application of GANs for time series, WaveGAN (Donahue *et al.* 2019) trains GANs by spectrogram images converted from audio data as time series for synthetic audio data generation. For financial time series, QuantGAN (Wiese *et al.* 2020) employs temporal convolutional networks (TCNs) for both its generator and discriminator. Furthermore TAGAN and TTGAN (Fu *et al.* 2022) introduce attentions and transformers into GANs and are compared to QuantGAN.

As another approach, variational autoencoders (VAEs) (Kingma and Welling 2014) are employed for financial time series generation (Dogariu *et al.* 2022). A VAE model has latent space and two neural networks, an encoder, and a decoder. The encoder is trained to embed given real data into latent space through the parameters of parametric models, and the decoder is trained to reproduce the given real data. Despite attempts using these generative models like GANs and VAEs to replicate financial time series, no model has yet fully captured all the stylized facts (Dogariu *et al.* 2022).

The characteristics denoted as stylized facts complicate the generation of synthetic financial time series. The related works in this area have generally focused on the replication of stylized facts (Cont 2001, Chakraborti *et al.* 2007, Ratliff-Crain *et al.* 2023), including fat tails, volatility clustering, seasonality, and calendar effects. All are absent in Brownian motions. As far as we know, the financial time series focused on in the related works are limited to stock prices, while other time series observed simultaneously with stock prices are out of scope.

In addition to GANs and VAEs, diffusion models are often used to generate images. The critical characteristics of generative models for images are quality, diversity, and the generation speed. The GAN, VAEs, and diffusion models satisfy two of these characteristics, although the third remains challenging. GANs offer quality and speed benefits, VAEs provide strength in diversity and speed, and diffusion models have quality and diversity advantages (Xiao *et al.* 2022). For applications to time series data, diffusion models are used for time series imputation (Tashiro *et al.* 2021).

3. Methodology

As discussed above, building a generative model to generate realistic financial time series remains a challenging task. Our goal is to leverage the strengths of DDPMs, originally developed for high-quality image generation, to synthesize realistic multivariate financial time series. We first convert the time series into images via wavelet transformation, then train a DDPM to generate new images, and finally invert the wavelet transformation to obtain synthetic time series data. Below, we provide a high-level summary of how DDPMs operate, followed by our wavelet-based preprocessing and postprocessing steps. In the subsequent description, we assume the simultaneous generation of three synchronized observed financial time series: prices, bid/ask spreads, and trading volumes.

3.1. Overview of denoising diffusion probabilistic models

Denoising Diffusion Probabilistic Models (DDPMs) (Ho *et al.* 2020) are a class of generative models that utilize a forward and reverse diffusion process to generate high-quality samples that approximate a complex data distribution. This section outlines the core formulation of DDPMs, with a focus on the underlying probabilistic framework. DDPMs operate based on a two-step Markovian process: the forward process (also known as the diffusion process) and the reverse process (also known as the denoising process). Given a data sample $x_0 \sim q(x)$, a predefined noise schedule $\beta_1, \beta_2, \ldots, \beta_T$ is employed to progressively corrupt the sample by adding Gaussian noise over T steps. This process is defined as:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I),$$

where $\mathcal{N}(x; \mu, \Sigma)$ denotes a normal distribution with a random variable x, a mean vector μ, and a covariance matrix Σ. By applying the reparameterization trick, the noisy sample x_t at any time step t can be explicitly computed as:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon,$$

where $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, and $\epsilon \sim \mathcal{N}(0, I)$. The reverse process aims to iteratively denoise $x_T \sim \mathcal{N}(0, I)$, ultimately recovering the clean sample x_0. It is parameterized by a deep neural network p_θ and follows:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)).$$

In practice, the model is trained to predict the noise ϵ added at each step by minimizing the following loss function:

$$\mathcal{L}(\theta) = \mathbb{E}_{x_0,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|^2\right].$$

This objective guides the network to learn an accurate reverse process that iteratively eliminating noise from a Gaussian prior to generate realistic data samples. After T steps, x_0 is a newly generated sample from the model. Figure 3 depicts this procedure schematically: moving from left to right illustrates the forward noising of real data, and moving from right to left illustrates the denoising that the model learns. This approach allows DDPMs to produce high-fidelity and diverse samples, as each small step in the reverse process is easier to learn than a single global mapping from noise to data.

3.2. Wavelet transformation for time series imaging

Preprocessing of financial time series: Our methodology for transforming financial time series into synthetic data unfolds through a series of designed steps. Initially, because sequence data with size 2^n simplify the subsequent discrete wavelet transformation, we expand the time series by mirror expansions at both ends to align the length of the time series to 2^n. Then we calculate the log returns of the stock prices by determining the differences of the natural logarithm of consecutive stock prices. This step, which is crucial for addressing the non-stationarity inherent in stock price time series, renders them more amenable to analysis. Concurrently, we apply the *arsinh* transformation to the trading volumes to approximate a logarithmic scale for large values by and to facilitate effective scale transformation. Unlike the natural logarithm, the *arsinh* function allows small values near 0 and 0 to enter (Bellemare and Wichman 2020). Following this, we undertake a power transformation on the series and normalize it as $\frac{(X_t - \mu(X_t))^{\frac{1}{p}}}{\sigma(X_t)}$. The power index can be different for each dimension in multivariate time series. In the steps so far, time series X_t contains many outliers, which reduce the training and inference efficiency. We substitute such outliers with the given z-values of the normalized data through winsorization. For example, if $X_t > z$ (resp. $X_t < -z$), then such X_t are substituted by z (resp. $-z$).

Wavelet transformation and filling of pixels of an image: After applying discrete wavelet transformation to the preprocessed time series, we obtain sequences of wavelet coefficients. For the transformation, we employed the Haar wavelet as the mother wavelet. The Haar wavelet is the simplest mother wavelet that provides a lossless transformation. This enables the inverse transformation from transformed coefficients to the original time series in the discrete wavelet transformation. Due to its simple formulation, the computational cost of the Haar wavelet is lower than that of other mother wavelets. Furthermore, its ability to detect edges of the original time series due to the step-function nature of the mother wavelet is suitable for analyzing time series with abrupt change points like financial time series, which exhibit properties known as stylized facts.

Because of the mirror expansion for making the length of the time series 2^n, we obtain a zero-th order coefficient, a first order coefficient, two second order coefficients, four third order coefficients, and so on up to $(n-1)$th order coefficients. These coefficients are regarded as luminance and are arranged to fill the pixels, transforming each time series into a monochrome image. In this pixel filling, kth order coefficients are embedded in the kth row of pixels. When the size of the coefficients of each order is 2^l, the row of pixels is split into 2^l areas, each of which is filled with the same coefficient value (figure 2). By treating these monochrome images as luminance channels for the three primary colors (red, green, and blue), we synthesize a color image from the trio of monochrome images. The synthesized color image represents a wavelet-transformed time series of price log returns, spreads, and trading volumes. This conversion from three time series data in one interval to a single color image is applied to multiple intervals, resulting in multiple color images as the training dataset. The culmination of this process involves the

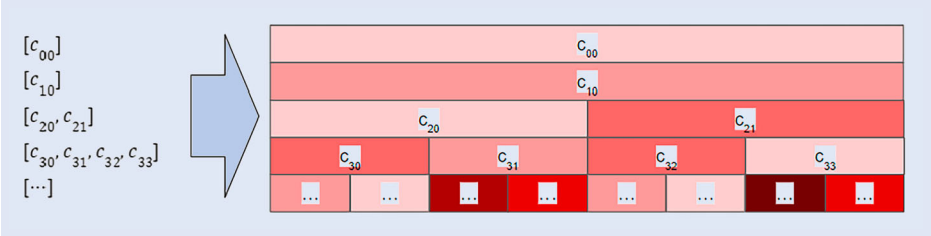


Figure 2. Pixel imaging. Wavelet coefficients $c_{00}, c_{10}, c_{20}, c_{21}, c_{30}, c_{31}, c_{32}, c_{33}, \ldots$ are tiled to image pixels.

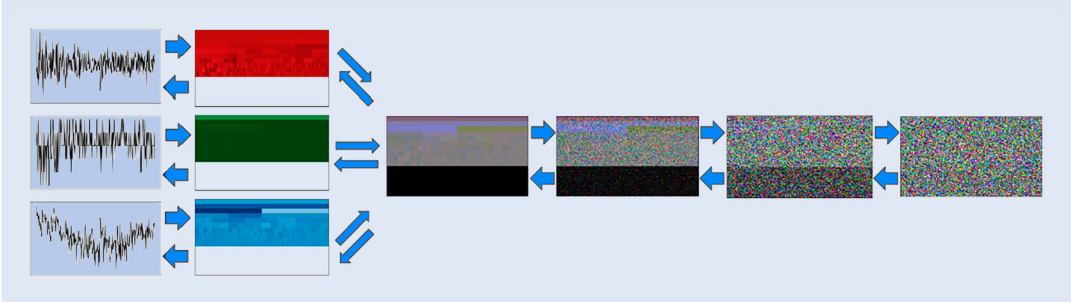


Figure 3. Overview of the methodology. (Left to right) We first convert the time series of log returns (top), spreads (middle), and trading volumes (bottom) into individual sets of wavelet coefficients, which are then mapped to color channels (RGB). The forward diffusion process progressively corrupts a real image with noise over multiple steps. (Right to left) The reverse diffusion process (i.e. the DDPM's learned denoising) iteratively recovers a synthetic image from pure noise. We finally apply an inverse wavelet transform to the generated spectrogram image to obtain the synthetic time series.

training of DDPMs using these color images converted from time series data. The synthetic images produced by the trained model are then converted back into time series data, employing the reverse operations of the preceding steps. Figure 3 summarizes these procedures. Through this process, we demonstrate how our approach generates synthetic financial time series data by leveraging the capabilities of DDPMs in learning and reproducing the complex dynamics of financial markets.

Model setup: DDPMs utilize a UNet architecture to facilitate learning through convolutional processes. This UNet is composed of multi-stage convolutions that include an attention mechanism. For this implementation, we employ the identical channel dimension parameters as found in the Hugging Face tutorial on DDPMs (https://huggingface.co/docs/diffusers/tutorials/basic_training): 128-128-256-256-512. These parameters define the channel dimensions at various UNet stages, enabling efficient processing and feature extraction across different scales of the input data.

4. Results

4.1. Experiments

We train the diffusion model to output 2-dimensional, 3-channel images that are converted from the time series of price log returns, spreads, and trading volumes by wavelet transformation.

Data: In our experiment, we selected minute-based stock prices, spreads, and trading volumes of AAPL.O traded on NASDAQ from January 2005 to December 2014. We purchased the data from Refinitiv (LSEG), a financial data provider. The data used in this study consist of minute-level OHLC (Open, High, Low, Close) prices for both bids and asks as well as minute-level trading volumes. During this period, some trading days have minutes without any trades, and we omit such entire trading days in these cases, resulting in 2481 sample days within this period. Because the liquidity of AAPL.O stock is sufficiently high, 2481 days of data are still available after this omission. One business day opens at 9:30 and closes at 16:00 EST, and we focus on the granularity of these 390 minutes within each trading day. Because the open and close times are fixed during this period, the lengths of one-day time series are identical across the samples. After mirror expansion in preprocessing, the length of the time series becomes 512. Through wavelet transformation and imaging the three time series, the price log returns, the spreads, and the trading volumes, these time series in one day become 16x256 images with three channels.

Parameters: In preprocessing, we adopt power conversion parameter $p = 1.5$ for price log returns and $p = 1.0$ for the spreads and trading volumes. The winsorization level is set at 10.0, meaning that outliers beyond 10σ are replaced by 10σ. The baseline logic is identical to the tutorial implementations on Hugging Face (https://huggingface.co/docs/diffusers/tutorials/basic_training). We train the model for 100 epochs, and other parameters follow the tutorial for DDPMs on Hugging Face.

Computational time: Our approach based on DDPM and wavelet transformation takes two hours for training under the above settings and another two hours to generate 2500

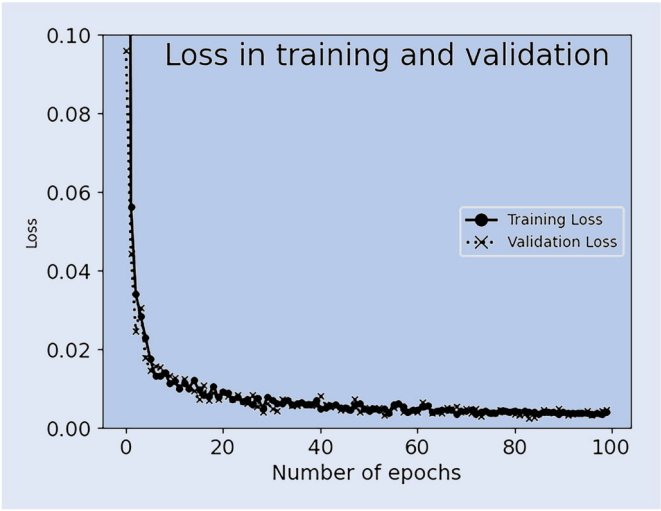


Figure 4. Losses in training and validation dataset.

images on an NVIDIA GeForce RTX 4090 using the PyTorch framework.

Comparison to other methodologies: Our methodology compares the outcomes of our DDPMs against established generative models for time series, such as TimeGAN (Yoon *et al.* 2019) and QuantGAN (Wiese *et al.* 2020). Regarding the DDPM-based approach, we compared it with discrete wavelet transformation and a simple approach, which fills 1x512 matrices with time series data after preprocessing and regards the three 1x512 matrices as one 1x512 color image. The comparison examines the essential stylized facts of financial markets, including their ability to replicate fat-tailed distributions, the slow decay of the autocorrelations of time series of volatilities, spreads and trading volumes, and the characteristic U-shaped pattern of intraday time series. In addition to these examinations, we also investigated the cross correlation coefficients among time series.

4.2. Evaluation

Losses in training and validation: Due to the nature of machine learning models, the values of a loss function decrease with regard to the number of epochs. If the values converge, the model's training works well. In this evaluation, we shuffled the original samples with 2481 days. 2000 days of data are regarded as the training dataset and 481 days of data are the validation dataset. Figure 4 explains the convergence of the loss function of the DDPM and the training is processed well.

Shape of time series: The comparison among TimeGAN, QuantGAN, and DDPMs (both with/without wavelet) reveals that TimeGAN's synthetic time series fail to convincingly replicate the nuanced movements of the log returns of the stock prices, a deficiency that was not rectified by such parameter adjustments as altering the dimensionality of the latent space (figure 5). This observation convinced us to focus our comparative analysis on QuantGAN and our DDPM-based approach in terms of log returns. To the best of our knowledge, no prior studies have generated time series data on spreads and trading volumes.

Fat-tailed distribution: The fat-tailed distribution analysis, illustrated in the probability density functions of the log returns (figure 6), demonstrates our approach's superiority when mirroring real distributions up to 10σ. QuantGAN, which uses a 'gaussianizer' in preprocessing to convert the original log return distribution to a Gaussian, exhibits poorer fit near 2σ of the distribution compared to the real data, although it fits well around 10σ. Our approach, based on DDPM and wavelet transformation, also shows good fits up to 10σ for the spreads and trading volumes.

Slow decay of autocorrelation: Some stylized facts pertain to the features of autocorrelations. In a previous work Cont (2001), the fast decay of autocorrelations is cited as a feature of log returns, and volatility clustering, which quantifies that high-volatility situations tend to cluster in time, is regarded as positive autocorrelations of volatilities and their slow decay. Here we show that our approach using DDPM and wavelet transformation explains both the fast decay of autocorrelations in the log returns and the positive autocorrelations with the slow decay in volatilities (figure 7). In minute-based time series, we also expect similar positive autocorrelation structures and their slow decays in the spreads and the trading volumes. Our approach also explains the existence of such positive autocorrelations and their slow decays.

Intraday seasonality: Furthermore, our analysis of intraday seasonality, which captures the patterns of the volatilities, the spreads, and the trading volumes, confirms the U-shaped pattern within a day. Figure 8 represents the average taken every minute for the given time series in the real data and for the generated time series. This pattern, starting with high values at the market's opening, dipping mid-day, and rising towards the close, is replicated by our approach, affirming its capacity to accurately mimic real market behaviors. QuantGAN and the simple approach using DDPM without wavelet transformation show flat patterns with regard to time, highlighting their limitations when representing intraday seasonality observed in real markets.

Cross correlation coefficients among time series: Because our data are synchronized time series observed within a day, we expect them to exhibit cross correlations. Table 1 is the cross correlation matrix among time series of real data. A positive correlation coefficient between volatility time series and trading volume time series indicates that high volatility and high trading volume tend to be observed simultaneously. A negative correlation coefficient, between spread time series and trading volume time series, suggests that market participants actively trade higher volumes under tight spreads. Table 2 represents the cross correlation matrix among time series generated by DDPM with wavelet imaging, and Table 3 shows the cross correlation matrix among time series by simple DDPM without wavelet imaging. Both DDPM-based approaches successfully replicate cross correlation coefficients among time series observed in real data: a positive correlation coefficient between volatilities and trading volumes, a negative correlation coefficient between the spreads and the

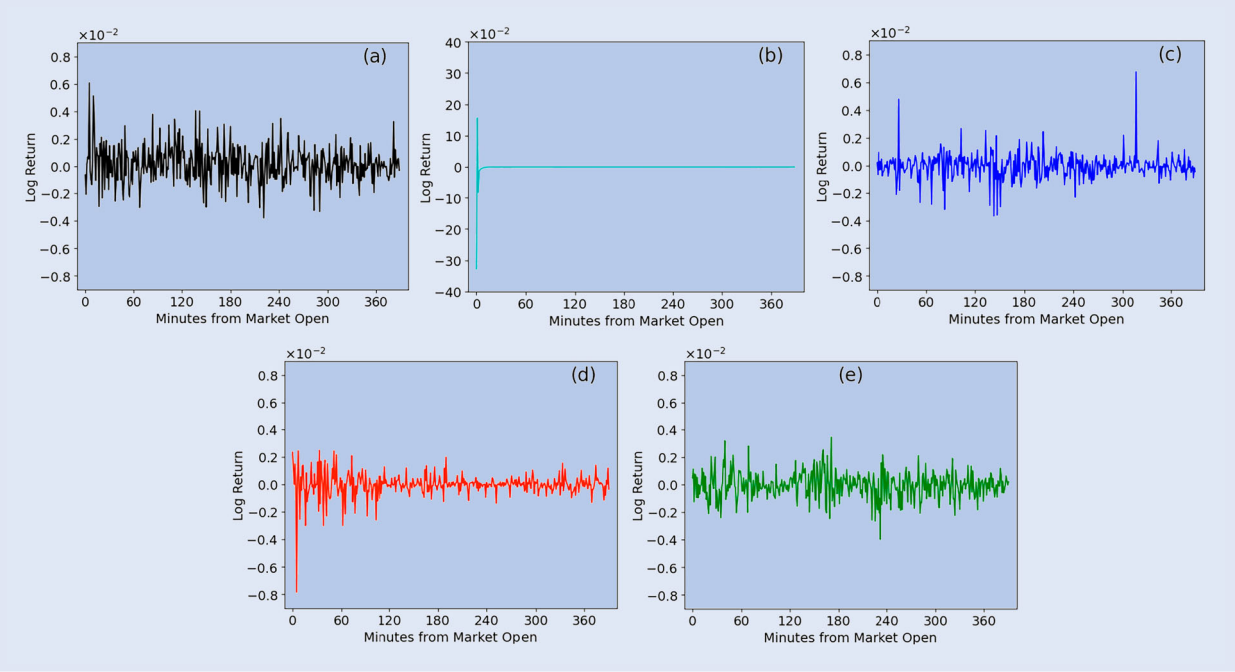


Figure 5. Shapes of time series of log returns. (a) real data as of January 7, 2005, (b) synthetic data by TimeGAN, (c) synthetic data by QuantGAN, (d) synthetic data by DDPM with wavelet imaging, and (e) synthetic data by DDPM without wavelet.

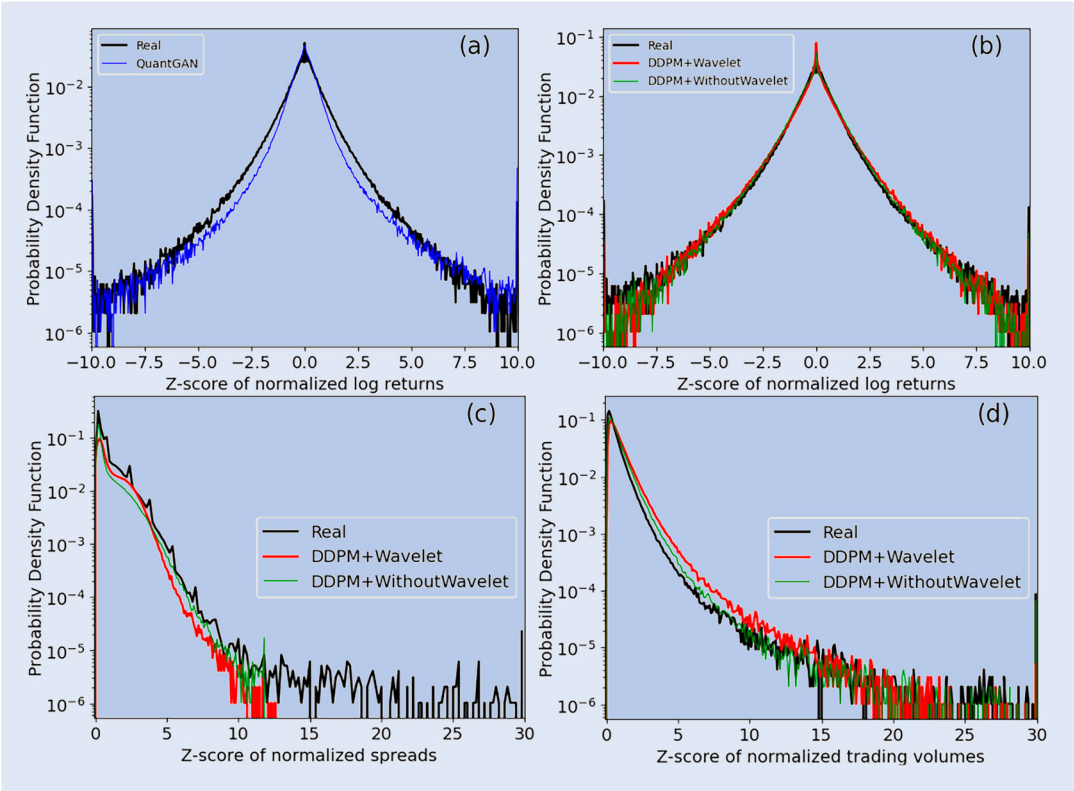


Figure 6. Probability density functions. (a) log returns of real data (black) and QuantGAN (blue), (b) log returns of real data (black), DDPM with wavelet imaging (red), and DDPM without wavelet (green), (c) spreads of real data (black), DDPM with wavelet imaging (red), and DDPM without wavelet (green), and (d) trading volumes of real data (black), DDPM with wavelet imaging (red), and DDPM without wavelet (green).

trading volumes, another negative correlation coefficient between volatilities and spreads, and other quite small correlations.

These findings demonstrate the robustness of our DDPM-based method for generating synthetic financial time series that faithfully reproduce complex market dynamics and

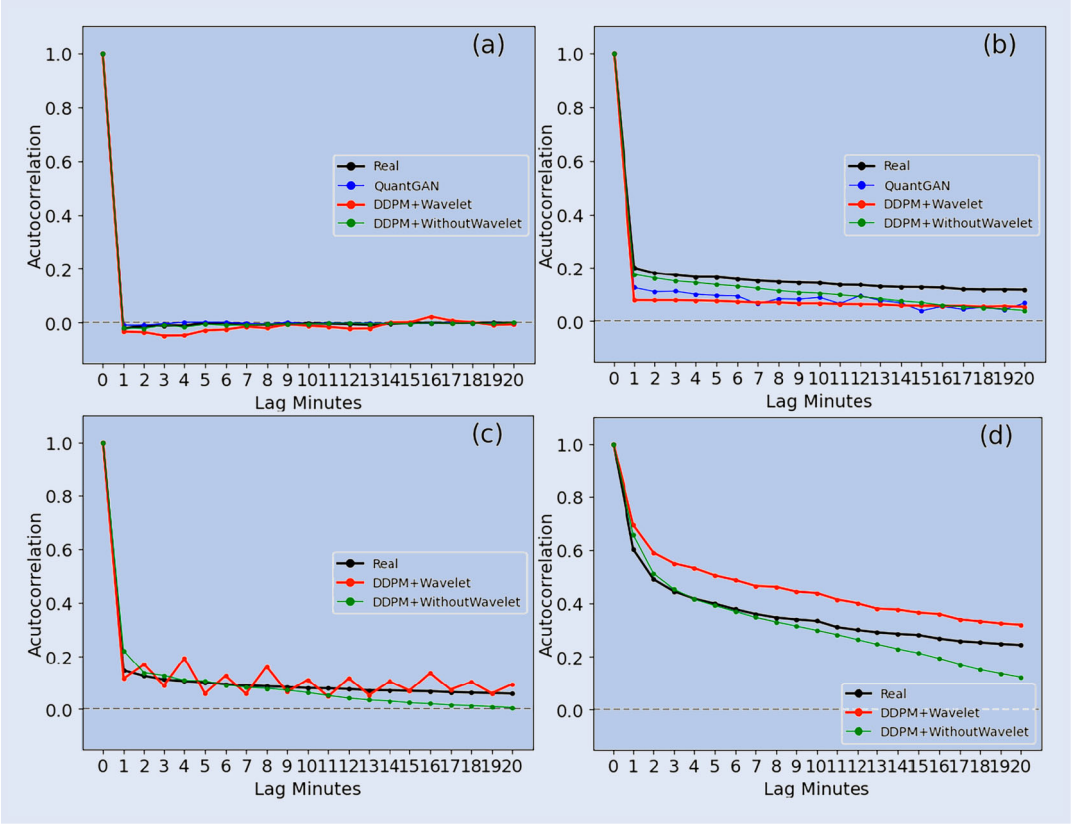


Figure 7. Autocorrelations. (a) log returns, (b) volatilities, absolute values of log returns, (c) spreads, and (d) trading volumes. In each chart, black chart represents real data, blue chart represents QuantGAN, red chart represents DDPM with wavelet imaging, and green chart represents DDPM without wavelet.

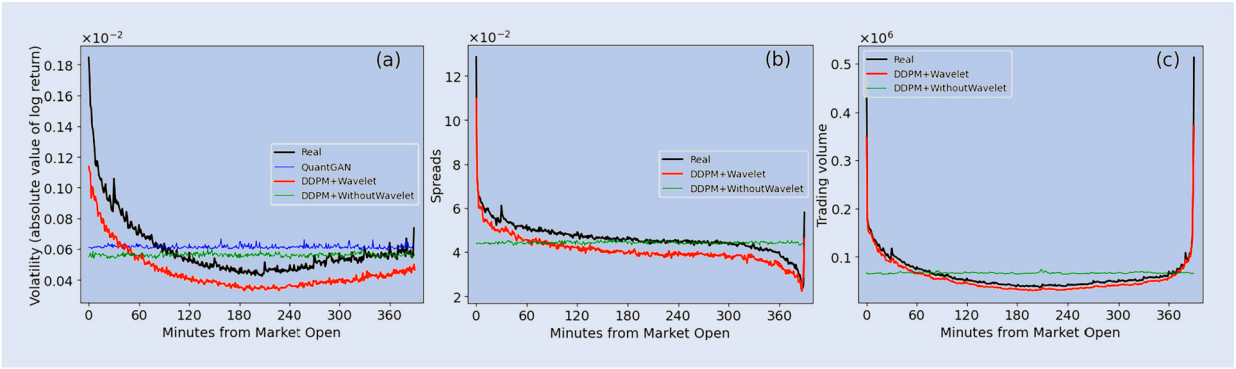


Figure 8. Intraday seasonality as average of per-minute time series over samples. (a) volatilities, absolute value of log return, (b) spreads, and (c) trading volume. In each chart, black chart represents real data, blue chart represents QuantGAN, red chart represents DDPM with wavelet imaging, and green chart represents DDPM without wavelet.

Table 1. Cross correlations among time series of real data.

	Log Returns	Volatilities	Spreads	Trading Volumes
Log Returns	1	−0.02	0.00	−0.02
Volatilities		1	−0.05	0.44
Spreads			1	−0.13
Trading Volumes				1

stylized facts. These check points are summarized in Table 4. Because TimeGAN failed to replicate the nuanced movements of the log returns of stock prices, we did not check other points. QuantGAN generated only the log returns of the stock prices; the cross correlations among multiple time series were skipped.

Table 2. Cross correlations among time series of synthetic data (DDPM+Wavelet).

	Log Returns	Volatilities	Spreads	Trading Volumes
Log Returns	1	0.00	0.01	0.00
Volatilities		1	−0.05	0.25
Spreads			1	−0.12
Trading Volumes				1

Table 3. Cross correlations among time series of synthetic data (DDPM+WithoutWavelet).

	Log Returns	Volatilities	Spreads	Trading Volumes
Log Returns	1	−0.01	0.00	0.00
Volatilities		1	−0.05	0.39
Spreads			1	−0.14
Trading Volumes				1

Table 4. Summary of comparisons among approaches.

	TimeGAN	QuantGAN	DDPM (without wavelet)	DDPM+Wavelet
Shape of time series	NG	OK	OK	OK
Fat tail	–	OK	OK	OK
Slow decays of autocorrelation	–	OK	OK	OK
Intraday seasonality pattern	–	NG	NG	OK
Cross correlation function	–	–	OK	OK

5. Conclusions

We suggested an alternative approach to generate synthetic time series by wavelet transformation and denoising diffusion probabilistic model (DDPM). The DDPM and wavelet imaging approaches more effectively replicated characteristics commonly observed in financial time series, such as the fat tails, the slow decay of autocorrelations including volatility clustering, the intraday seasonality patterns, and the cross correlation coefficients among time series, compared to TimeGAN and QuantGAN approaches and the simple application of DDPM without wavelet imaging. The DDPM and wavelet approach especially had a better representation of intraday seasonality in the actual market data than these methodologies in comparison. Imaging through the wavelet transformation can capture intraday seasonality because the relationship among frequencies is more explicit than the original time series. In the preprocessing of imaging, the DDPM with a wavelet imaging approach fills the top rows of an image with a zero-th wavelet coefficient, which represents the overall intraday trend. The next and subsequent pixel rows of the image represent progressively finer market microstructures. As a result, the information of short-term microstructures with multiple time scales by wavelet transformation around market open (resp. close) is embedded in the left (resp. right) side of the image. This contributes to the representation of the U-shape structures of the observed intraday data.

In conclusion, our application of denoising diffusion probabilistic models (DDPMs) in conjunction with wavelet image transformation was an effective method for generating synthetic financial time series that closely adhere to underlying stylized facts. The strategic use of RGB channels in color images to represent and simultaneously generate three interconnected time series replicated the structure of the cross correlation functions among the multiple time series, representing a significant advancement in the field. Looking ahead, the potential to extend this methodology to utilize three or more channels might allow simultaneous generation of multiple correlated stock prices. The methodology is generalizable across different assets including those beyond AAPL.O stock. However illiquid stocks often have minutes with no trades on more days compared to AAPL.O. In this case simply omitting entire trading days as we do for AAPL.O could ignore characteristic features of illiquid stocks. In this case, different preprocessing of the original time series, for example, step interpolation of mid prices and spreads, and zero trading volume during the minutes without any trades, could yield better results. In addition, finding an alternative mother wavelet that outperforms the Haar wavelet could be another avenue for future research. Building on this foundation, future work is poised to explore the generation of synthetic data that capture even more nuanced relationships and cross correlations within financial markets, extending the boundaries of what is possible in synthetic data generation and financial modeling.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Japan Science and Technology Agency CREST under Grant JPMJCR20D3; and

Japan Society for the Promotion of Science under Grant JP23K21018.

Data availability statement

In this study, we used minute-based stock prices, spreads, and trading volumes of AAPL.O traded on NASDAQ from January 2005 to December 2014. This data is available from Refinitiv (LSEG), a financial data provider. We do not have any special access privileges to this database. Other researchers can access the data by obtaining a license contract with Refinitiv through their information provision service at https://www.lseg.com/.

ORCID

Tomonori Takahashi http://orcid.org/0009-0004-7652-4019
Takayuki Mizuno http://orcid.org/0000-0003-0673-2707

References

Bachelier, L., Théorie de la spéculation. *Ann. Sci. l'E.N.S. (SMF)*, 1900, **17**, 21–86.
Bellemare, M.F. and Wichman, C.J., Elasticities and the inverse hyperbolic sine transformation. *Oxf. Bull. Econ. Stat.*, 2020, **82**(1), 50–61.
Bollerslev, T., Chou, R.Y. and Kroner, K.F., ARCH modeling in finance: A review of the theory and empirical evidence. *J. Econom.*, 1992, **52**(1-2), 5–59.
Brophy, E., Wang, Z., She, Q. and Ward, T., Generative adversarial networks in time series: A systematic literature review. *ACM Comput. Surv.*, 2023, **55**(10), 199, 1–31.
Chakraborti, A., Patriarca, M. and Santhanam, M.S., Financial time-series analysis: A brief overview. In *Econophysics of Markets and Business Networks*, edited by A. Chatterjee and B.K. Chakrabarti, pp. 51–67, 2007 (Springer: Milan).
Cont, R., Empirical properties of asset returns: Stylized facts and statistical issues. *Quant. Finance*, 2001, **1**(2), 223–236.
Das, A., Kong, W., Sen, R. and Zhou, Y., A decoder-only foundation model for time-series forecasting, 2023. Available online at: https://arxiv.org/abs/2310.10688 (accessed 27 September 2024).
Dogariu, M., Ştefan, L.D., Boteanu, B.A., Lamba, C., Kim, B. and Ionescu, B., Generation of realistic synthetic financial time-series. *ACM Trans. Multimedia Comput. Commun. Appl.*, 2022, **18**(4), 96, 1–27.
Donahue, C., McAuley, J. and Puckette, M., Adversarial audio synthesis. Paper presented at International Conference on Learning Representation (ICLR 2019).
Eckerli, F. and Osterrieder, J., Generative adversarial networks in finance: An overview, 2021. Available online at: https://papers.ssrn.com/abstract=3864965 (accessed 27 September 2024).
Gabaix, X., Power laws in economics and finance. *Annu. Rev. Econom.*, 2009, **1**, 255–294.
Gabaix, X., Gopikrishnan, P., Plerou, V. and Stanley, H.E., A theory of power-law distributions in financial market fluctuations. *Nature*, 2003, **423**(6937), 267–270.
Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y., Generative adversarial nets. Paper presented at the International Conference on Neural Information Processing Systems (NIPS 2014), 2672–2680, 2014.
Ho, J., Jain, A. and Abbeel, P., Denoising diffusion probabilistic models. Paper presented at the 34th International Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada, 2020.
Kingma, D.P. and Welling, M., Auto-encoding variational Bayes. Paper presented at the International Conference on Learning Representation (ICLR 2014), 2014.
Lux, T., Stochastic behavioral asset-pricing models and the stylized facts. In *Handbook of Financial Markets: Dynamics and Evolution*, edited by T. Hens and K.R. Schenk-Hoppe, pp. 161–215, 2009 (Science Direct).
Mizuno, T., Kurihara, S., Takayasu, M. and Takayasu, H., Analysis of high-resolution foreign exchange data of USD-JPY for 13 years. *Physica. A.*, 2003, **324**(1-2), 296–302.
Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L.A. and Stanley, H.E., Econophysics: Financial time series from a statistical physics point of view. *Physica. A.*, 2000, **279**(1-4), 443–456.
Ramsey, J.B., Usikov, D. and Zaslavsky, G.M., An analysis of U.S. stock price behavior using wavelets. *Fractals*, 1995, **03**(02), 377–389.
Ratliff-Crain, E., Van Oort, C.M., Bagrow, J., Koehler, M.T. and Tivnan, B.F., Revisiting stylized facts for modern stock markets. Paper presented at 2023 IEEE International Conference on Big Data (BigData), Sorrento, Italy, 15–18 December, 2023.
Samanidou, E., Zschischang, E., Stauffer, D. and Lux, T., Agent-based models of financial markets. *Rep. Prog. Phys.*, 2007, **70**(3), 00.
Shakeel, M. and Srivastava, B., Stylized facts of high-frequency financial time series data. *Global Bus. Rev.*, 2018, **22**(2), 550–564.
Takayasu, M., Watanabe, K., Mizuno, T. and Takayasu, H., Theoretical base of the PUCK-model with application to foreign exchange markets. In *Econophysics Approaches to Large-Scale Business Data and Financial Crisis*, edited by M. Takayasu, T. Watanabe and H. Takayasu, pp. 79–98, 2010 (Springer: Tokyo).
Tashiro, Y., Song, J., Song, Y. and Ermon, S., CSDI: Conditional score-based diffusion models for probabilistic time series imputation. Paper presented at the 35th Conference on Neural Information Processing Systems (NeurIPS 2021), 2021.
Wiese, M., Knobloch, R., Korn, R. and Kretschmer, P., Quant GANs: Deep generation of financial time series. *Quant. Finance*, 2020, **20**(9), 1419–1440.
Fu, W., Hirsa, A. and Osterrieder, J., Simulating financial time series using attention, 2022, Available online at: https://arxiv.org/abs/2207.00493 (accessed 2 March 2025).
Xiao, Z., Kreis, K. and Vahdat, A., Tackling the generative learning trilemma with denoising diffusion GANs. Paper presented at the International Conference on Learning Representations (ICLR 2022), 25–29, April 2022.
Yoon, J., Jarrett, D. and van der Schaar, M., Time-series generative adversarial networks. Paper presented at the 33rd International Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada, 2019.