A Methodology for Validating Diversity in Synthetic Time Series Generation

Fouad Bahrpeyma[a,*], Mark Roantree[c], Paolo Cappellari[b], Michael Scriney[a], Andrew McCarren[a]

[a]Insight Centre for Data Analytics, School of Computing, Dublin City University, Dublin 9, Ireland.
[b]City University of New York, 2800 Victory Blvd, Staten Island, 10314 NY, USA.
[c]VistaMilk SFI Research Centre, Dublin City University, Dublin 9, Ireland.

Abstract

In order for researchers to deliver robust evaluations of time series models, it often requires high volumes of data to ensure the appropriate level of rigor in testing. However, for many researchers, the lack of time series presents a barrier to a deeper evaluation. While researchers have developed and used synthetic datasets, the development of this data requires a methodological approach to testing the entire dataset against a set of metrics which capture the diversity of the dataset. Unless researchers are confident that their test datasets display a broad set of time series characteristics, it may favor one type of predictive model over another. This can have the effect of undermining the evaluation of new predictive methods. In this paper, we present a new approach to generating and evaluating a high number of time series data. The construction algorithm and validation framework are described in detail, together with an analysis of the level of diversity present in the synthetic dataset.

Keywords: Synthetic Time Series, Time Series Features, Time Series

Metrics, Diversity, Coverage.

Table 1: Specification Table

Subject Area:	Data Science
More specific subject area:	Time Series
Method name:	A Methodology for Validating Diversity in Synthetic Time Series Generation.
Name and reference of original method:	N.A
Resource availability:	Method & Datasets available on Zenodo.

Introduction

Time series analysis has long been an interesting topic of research across multiple domains, as many systems require a sequential monitoring of their data streams at constant intervals. Daily prices, weekly stock indices, hourly temperature and monthly precipitation rates are examples of the domains where a sequential monitoring is incorporated. There are currently multiple challenges in time series analysis such has time series classification, time series clustering, feature learning and time series prediction. Time series classification is a process employed to label unseen time series into a set of pre-existing classes of time series [1]. Time series classification has applications in various domains such as EEG signal analysis and fault detection [1]. Feature learning is the process of extracting and learning features from time series data in order to improve time series classification. Many studies have been conducted in feature learning such as [2] and [3] which generate discriminative time series features in an attempt to improve the performance of time series classification.

Our previous studies in time series prediction [4] suggest that evaluating new methods for time series predictions generally takes place using datasets that are specific to a researcher's area of interest. The acquisition of data involves harvesting from specific studies or bodies that have an interest in a particular research question. However, this can mean that solutions are not applicable to other domains. There are numerous competition and open

source datasets repositories that have been made available to the research community, such as BCI [5] and Kaggle [6]. These datasets have allowed researchers to test their methods more extensively, but the datasets are not typically classified by *time series* evaluation metrics but by *accuracy* of the methods employed.

The primary issue for time series researchers is the lack of available data to ensure robust validation. Synthetic data has been mixed with real life data in an attempt to broaden the scope of existing methods but the range of time series characteristics is quite narrow and datasets can be relatively small. There are many examples of researchers generating synthetic time series, such as: synthetic data generation [7]; surrogate data analysis [8]; using heuristics to materialize datasets [9]; and simulated data [10]. Surrogate data analysis can be used to estimate the impact of the *scale* of a time series characteristic through the comparison of the given time series with surrogate series [8]. This can then be used to estimate the impact of non-linearity in a time series in comparison to a series generated from a linear models such as ARIMA [11] and thus, allows researchers to replicate statistical features such as auto-correlation.

Time series clustering is a type of analysis that identifies similar time series and places them into a set of distinct groups. Time series clustering has applications in domains such as community detection and social media analysis. Studies such as [12] present new features known as *shapelets* to improve time series clustering. Shapelets are useful as features for classification and clustering but as it is not a statistical time series feature, it cannot be measured quantitatively to demonstrate diversity and coverage. Other approaches included [7], where the authors used a Markov chain model to create synthetic time series and an approach synthetic data creation for specific types of time series such as critical transitions [13]. In [14], a stochastic approach was presented to reproduce *long range* persistence of time series at multiple scales. An approach called Generative Adversarial Networks (GAN) [15] adopted machine learning algorithms to generate time series.

Problem Definition. Time series data generally exhibits a number of key characteristics (described later) and an important consideration when generating synthetic data is for these properties to occur with a fairly even level of distribution. We refer to this as *diversity* of the time series. Data generation for simulating changing environments was studied in [16] but no benchmark was developed for these simulations. In effect, they did not base their study on the fundamental structure of time series data and its rela-

tionship with time series characteristics. In fact, none of these studies have investigated the presence of differing characteristics in the generation of time series. Instead, we will show that they focus on the creation of either visually similar series or the presence of one particular feature e.g. [14].

Contribution. In this paper, we present a methodology to create synthetic time series with the primary aim of ensuring diversity of time series characteristics across the overall dataset. By diversity, it implies that datasets were built to incorporate time series characteristics such as long and short term dependence, non-stationarity, kurtosis, skewness, trend and varying degrees of complexity. This is not the same as introducing uncertainty [17] which would instead mean that the dataset contains time series of various types and behaviors. Diversity is a desirable attribute especially for time series algorithms evaluation purposes, where the algorithms' performances are required to be evaluated against a wide range of possible situations. We also provide a rigorous validation to measure the degree to which each time series property is contained within the dataset. As part of this work, we generated 53,637 time series to be shared with the time series community [18]. To the best of our knowledge, no other study has constructed this volume of time series *together with a robust evaluation for diversity.* In terms of exploiting this resource, the dataset is provided in full, with evolving documentation to describe its usage together with a link to this paper to provide the researcher with an understanding of the characteristics of subsets of the time series. In terms of data provenance, this paper provides a detailed description of the algorithm used to generate the data. It is also anticipated that by tweaking parameter settings, this repository could grow to more significant number of time series and potentially grow or accelerate research in time series prediction.

Paper Structure. The remainder of this paper is organized as follows: in section 2 (The Need for Reliable Time Series Data), we motivate a requirement for this type of method and examine how and why other researchers have created synthetic time series data; in section 3 (Generating Time Series), our method for constructing synthetic time series is presented; in section 4 (A Feature Set to Capture Diversity), we present the fundamental features used to evaluate the synthetic time series; in section 5 (Evaluation), we present our validation together with a detailed discussion of the results; and in the final section, we present our conclusions.

4

The Need for Reliable Time Series Data

Traditional (statistical) applications of time series prediction have been practiced under the assumption that the time series was produced by a linear continuous process [19]. However, this may not be the case where time series are the output of interactions of many alternating series and thus, linearity cannot always be assumed [20]. However, many processes such as financial time series are fundamentally characterized by complex, substantially noisy, dynamic, and nonlinear behaviour [21].

Data generation has been widely used in time series analysis through the use of surrogate data analysis [8], synthetic data generation [7] or simulated data[10]. Surrogate data analysis can be used as a means of estimating the impact of the scale of a characteristic in a time series, through the comparison of the given time series with surrogate series [8]. This can be demonstrated by estimating for example, the impact of non-linearity in a time series in comparison to a series generated from a linear models such as ARIMA and thus, allows researchers to replicate statistical features such as auto-correlation [22].

The majority of practices in time series generation typically use linear approaches, such as the ARIMA family of models. These models establish fundamental statistical consistency, by means of reproducing the mean, variance and auto-correlations of lags of the parent historical data [7]. However, many real-world time series show substantially more complex statistical properties; for example, time series with skewness rather than Gaussian distributions, or those characterized by statistical inter-dependencies [23].

In [7], a Markov chain model was used to generate synthetic data for a wind speed time series analysis. Characteristics such as mean, standard deviation and frequency distribution were predominantly used as assessment metrics. They also evaluated auto-correlation and power spectral density to determined the persistence structure of the series.

In [24], the authors presented a method that incorporates maximum entropy bootstrap to generate ensembles for the given time series data. However, this method only focuses on the low frequency approximation of the signal and discards memory characteristics laid on temporal fluctuations. In [25], the authors also focused on the shape of the signal and tried to use white noise to generate new patterns. This work was originally conducted to compare the performances of time series classification methods on the data for variant representations. However, this work did not address the role of

diversity of time series in performance comparisons.

In [26], the authors present a similarity measure that studies generation methods for general time series features. Their work also presents a feature-based time series generation approach that evolves cross-domain time series datasets. The authors present a generic method capable of generating time series from a diversity of domains, as opposed to previous methods that generate time series for particular domains such as weather, economics and energy. This work introduced 4 general attributes for a time series generation method: Dataset-oriented, Deterministic, Stochastic and Innovative. The *Innovative* feature suggests an overlap with our work as it provides a reference to the requirement for *diversity*. However, unlike our approach, their research is bound to domain-specific constraints as it requires examples from the domain for which synthetic time series are to be generated.

In [27], the authors presented a method known as GRATIS and used Mixture AutoRegressive (MAR) models to generate time series data. They incorporated 26 time series features and used a genetic algorithm to evolve time series and create new instances. This approach generated 20,000 yearly, 20,000 quarterly, 40,000 monthly and 10,000 weekly time series based on the MAR models. This work compared their synthetic time series with those of M4 to provide an analysis of coverage and diversity, using M4 as the reference dataset. However, their measures indicate diversity and miscoverage only in relation to the reference dataset after dimension reduction, which is analytically difficult to project into the original feature space. We believe that our approach is free from this limitation.

In [14], a stochastic approach was presented to simulate long range persistence of hydrometeorological time series at multiple scales. The authors use a linear stochastic model to generate synthetic data that replicates the Hurst-Kolmogorov characteristics of the original process. However, this method attempts to replicate temporal dynamics to create similar series, and thus cannot produce diverse series.

More recently, a method called Generative Adversarial Networks (GAN) [15] received attention for generating similar datasets. GANs were originally introduced as an approach that facilitates generative modeling via deep learning. The GANs' training process is to force the output of the network to follow the distribution of the given input. Most of the studies on GANs focus on image generation and limited work address time series data. [28] was the first attempt that used GANs to generate continuous sequential data (which is a superset of time series). This work tries to generate new music pieces

based on some reference classical musics. A similar attempt was also made in [29], and used GANs to reproduce musical symbolic sequences. Past studies on GANs have also addressed diversity, such as in [30], although in GANs diversity has received a greater attention in the context of training performance, where diversity is required in training samples in order to stabilize modeling performance. More complex practices have also been reported that use deep learning for synthetic data generation. In [31], the authors proposed a deep learning architecture which incorporates a stack of multiple Long-Short-Term-Memory (LSTM) networks and a Mixture Density Network (MDN) for Synthetic Sensor Data Generation. This work attempted to develop a model that reproduces sequences of data which preserve specific statistical properties. However, this work did not consider the diversity of the synthetic data, a characteristic that we believe is crucial when validating results. In fact, none of these studies investigated the presence of differing characteristics when generating synthetic time series. Instead, they focus on the creation of either visually similar series or the presence of one particular feature, as in [14].

In summary, almost all research on surrogate and synthetic time series generation were conducted to reproduce the *same set* of features with *small* variations. To date, there has not been an extensive generation of time series datasets that cover a broad range of time series characteristics and such a method will facilitate a more robust validation of future time series models.

Generating Time Series

A time series is a sequence of equally spaced time ordered data points. When conducting time series analysis, the predominant objective is to understand the characteristics of the data, and the extraction of meaningful statistics. Time series data can be broken up into four components [32, 33]: *trend*, *seasonality*, *cyclicality*, and *irregularity*.

These components can be used to drive the generation of time series. It is important to note that there is no standard way of combining these components to generate a time series, especially when considering that each component can itself be generated in multiple ways. For instance in [34], the irregularity component is used as the base and manipulated by adding the trend and the seasonality components, which are combined in an additive or multiplicative way.

Time Series Components

Trend. The trend T_t describes the long term increase or decrease in the data. The trend can be linear or not, and can be described via the equation in Eq. 1.

$$T_t = (ax + b) \times sin(x) + cx + d \tag{1}$$

In Eq. 1, symbols a, b, c, d are the coefficients that allow one to specify the desired behavior of the trend, specifically: c, d control the linearity, while a, b via the sinusoidal function determine the non-linear behavior.

Seasonality and Cyclicality. Both seasonality and cyclicality describe repeating behaviors in the time series data. Specifically, seasonality describes a repeated behavior that occurs at regular intervals (e.g. every number of seconds, days, weeks, etc.); cyclicality, on the other hand, describes repeated behaviors that occurs at irregular intervals. We used four functional forms of repeating patterns to simulate seasonality S_t: the sinusoidal function, the step-wise function, the impulsive function and the triangular function.

The *Sinusoidal function* is simulated using Eq. 2, where, β_0 and β_1 are the weights; α_0 and α_1 are the phases for the sinusoidal functions; and $\alpha_0, \alpha_1, \beta_0$ and β_1 are constants.

$$S_t = \beta_0 sin(\alpha_0 t) + \beta_1 sin(\alpha_1 t) \tag{2}$$

The *Step-wise* function is a type of latch function (with two stable states) that changes between two values at fixed intervals, shown in equation 3, where, p $(p \geq 0)$ is the period, t is time and m is an integer.

$$S_t = \begin{cases} 1 & 2mp < t < 2mp + 1 \\ 0 & 2mp + 1 < t < 2mp + 2 \end{cases} \tag{3}$$
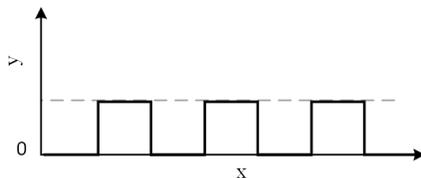
The step-wise function is shown in Figure 1.

Figure 1: The Step-wise function

The **Triangular** function is similar to the step-wise function where the step-wise effect occurs in the slope of the line. Equation 4 implements the triangular function with a fixed slope α, with b_0 and b_1 as constants, p as the period, t as time, and m as an integer.

$$y = \begin{cases} \alpha t + b_0 & 2mp < t < 2mp + 1 \\ -\alpha t + b_1 & 2mp + 1 < t < 2mp + 2 \end{cases} \tag{4}$$
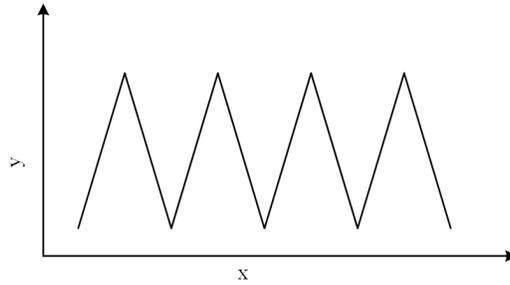
The Triangular function is shown in Figure 2.

Figure 2: The Triangular function

The *Impulsive function* is a pattern that has a value of 1 at fixed intervals and 0 otherwise and was implemented using 5, where t is time, and p is the period.

$$y = \begin{cases} 1 & \lfloor t/p \rfloor = t/p \\ 0 & otherwise \end{cases} \tag{5}$$

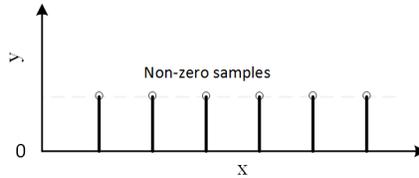The Impulsive function is shown in Figure 3.

Figure 3: The Impulsive function

Note that the impulsive function is smoothed using a moving average operation to avoid sharp fall and rise fluctuations.

9

Irregularity. The Irregular component describes behaviors that cannot be represented via the trend or the cyclicality/seasonality. In time series analysis this component is often referred to as *noise*. Some researchers, e.g. [35], believe that this component carries important information. Therefore, in order to accommodate for these theories, we model the noise as a signal with its own characteristics. Specifically, we consider following three models:

- Fractional Gaussian noise (fGn), which represents stationary series with a constant mean and variance;

- Fractional Brownian motions (fBm), which are non-stationary series with time-dependent variance [36];

- Multi-fractal Brownian motion, for the case where the Hürst exponent is applicable to time series, that is: the index is a function of time.

The Hurst exponent H [37] is one of the most popular methods to measure Long Range Dependence (LRD). H attempts to explain LRD as a property of stochastic self-similar processes. Here, $x(t)$ is self-similar with the Hurst exponent H, when for a stretching factor λ, the rescaled process $x(\lambda t)$ is equal to the original process $x(t)$ in terms of distribution as in Eq. 6, where $\doteq$ denotes *equality* in terms of distributions.

$$x(t) \doteq \lambda^{-H} x(\lambda t) \tag{6}$$

If the fluctuations are stationary (the process has a constant mean and a constant variance), the process is said to have fractional Brownian motion (fBm). Based on [38], the auto-correlation function for fBm processes is defined in Eq. 7.

$$\rho(k) = \frac{1}{2}(|k+1|^{2H} - 2|k|^{2H} + |k-1|^{2H}) \tag{7}$$

Based on [37], applying a first-order Taylor expansion to $\rho(k)$ from Eq. 7 delivers the functionality in Eq. 8, for $k \to \infty$.

$$\frac{\rho(k)}{H(2H-1)|k|^{2H-2}} \to 1 \tag{8}$$

It can be inferred from Eq. 8 that the autocorrelation $\rho(k) \propto |k|^{2-2H}$ when $H > \frac{1}{2}$, based on [37].

Our Combinational Methodology

Building on the work presented in [39], we combine *trend* and *cyclicality* into a joint component known as the *trend-cycle* component in order to prevent known complexities involved in identifying *Cyclicality*. In our time series construction method, we consider all possible additive and multiplicative combinations of *trend-cycle* T_t^c, *seasonality* S_t and *irregularity* I_t, using the approach presented in [40], where there are 8 possible models for combining T_t^c, S_t and I_t, shown in Table 2.

Table 2: Time Series Component Combinations

Model	Description
Model 1	$Y_t = T_t^c + S_t + I_t$
Model 2	$Y_t = (T_t^c + S_t)I_t$
Model 3	$Y_t = (T_t^c + I_t)S_t$
Model 4	$Y_t = (S_t + I_t)T_t^c$
Model 5	$Y_t = T_t^c S_t + I_t$
Model 6	$Y_t = T_t^c I_t + S_t$
Model 7	$Y_t = S_t I_t + T_t^c$
Model 8	$Y_t = T_t^c S_t I_t$

In Table 2, Model 1 is the pure additive model which is the most widely used model in the time series community. Model 8, or the pure multiplicative model, is the second most popular model among time series researchers. The other models in Table 2 are also used in the time series studies with *Model 3* and *Model 5* being more popular because they incorporate irregularity I_t using an addition operation.

An example of combining time series components is illustrated in Figure 4.

A Feature Set to Capture Diversity

In the previous section, we described a method that ensures diversity in synthetic time series using different combinations of functions that introduce time series characteristics. As one cannot directly measure time series components such as trend, it is necessary to extract a feature set to support any
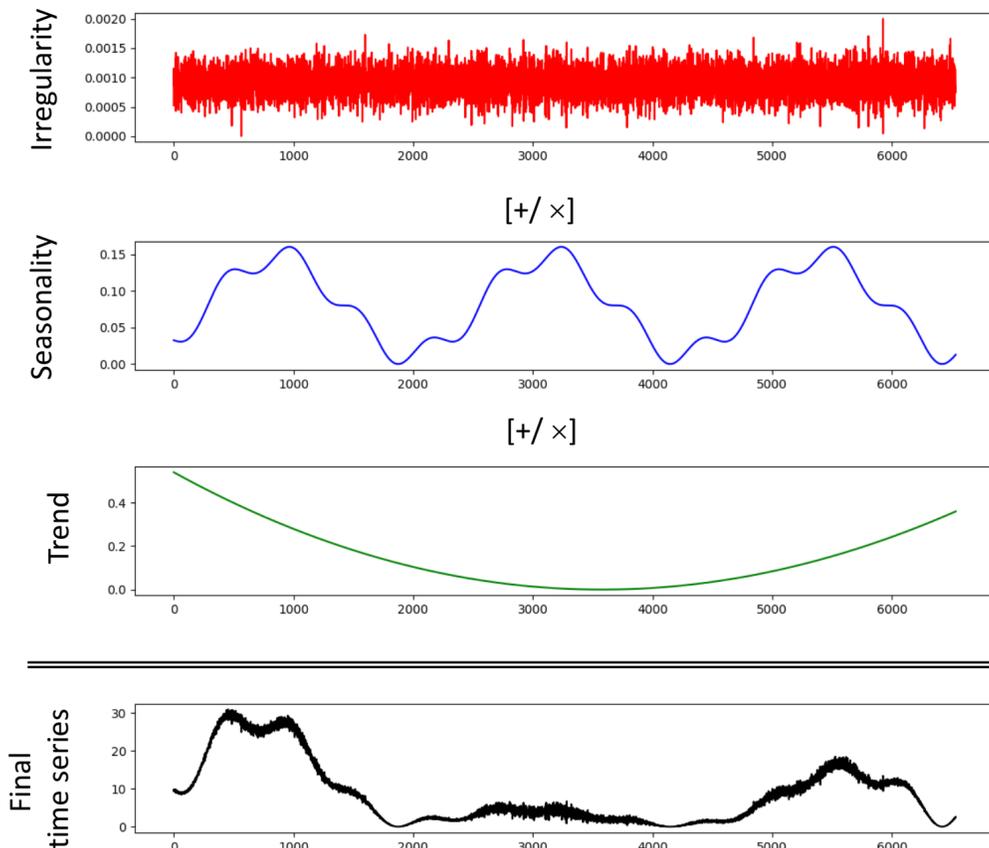
Figure 4: Combining Time Series components

evaluation. For our validation framework in this paper, we use *Long-Range Dependence*, *Complexity* and *Normality* .

LRD

Long-Range Dependence (LRD) measures the degree of dependence (correlation) over long intervals of time, which is a way to indicate "memory" in a time series. As mentioned earlier, the Hurst exponent is the traditional method for measuring LRD [41]. The Hurst exponent divides time series data into three categories: Negatively-correlated $0 < \alpha < 0.5$, Uncorrelated $\alpha \simeq 0.5$ and Correlated $0.5 < \alpha < 1$. However, the Hurst exponent is only able to process *stationary* time series. An alternative to the Hurst exponent for measuring LRD is the Detrended Fluctuation Analysis (DFA) approach

which allows for the detection of LRD in *non-stationary* time series.

Using Detrended Fluctuation Analysis (DFA), LRD can be assessed and categorized into 6 well known and critical classes.

Values of α can be interpreted as follows: $\alpha = 1$ indicates perfect (self) similarity in the data (a characteristic of the Self Organized Critically systems [42]); $\alpha = 1/2$ represents white noise, no similarity (or no memory); $1/2 < \alpha < 1$ describes positive correlation, with similarity (memory) increasing with the values of α; $\alpha < 1/2$ indicates inverse correlation; $\alpha > 1$ indicates that while correlations exist, they cannot be described in the form of a power-law relationship. A special case where $\alpha = 1.5$, indicates Brownian noise or the integration of white noise. α also provides information about the roughness of the time series where larger values of α belong to smoother time series. $1/f$ noise can be interpreted as a compromise between the complete unpredictability of white noise (very rough landscape) and the very smooth landscape of Brownian noise.

Complexity

Entropy can be used to measure complexity [43]. As per [44], given a signal y with sample size N and tolerance r, sample entropy is the negative logarithm of the conditional probability that a sub-series of length m matches point-wise with the next point with tolerance (distance less than) r. In this paper, we used *spectral* entropy to evaluate complexity, which measures the uniformity of the power spectrum distribution or the frequency component distribution.

Normality

Normality is a test to determine if data falls into a normal distribution [45]. Common metrics to measure normality are: Kurtosis, Skewness, and Gaussianity of the Differences (GoD). Kurtosis measures the number of outliers in the dataset with respect to a normal distribution: when Kurtosis is high, the dataset has a higher number of outliers (heavy tail in the distribution); when kurtosis is low, the outliers are low to none (light tail). Skewness measures the symmetry of the distribution: when positive, the distribution has a longer or fatter tail on the right side; when negative, the left side of the distribution has a longer or fatter tail; when zero, the distribution is symmetrical. Gaussianity of the differences (GoD) was used in [33, 46]

to measure the normality of the distribution of the changes (the first difference of the series) in the time series: this is an important metric because differencing is an important phase of many time series analysis.

Evaluation

As the goal of this research is to provide researchers with a method to create *diverse* time series datasets, it must be accompanied by validation framework to *measure* diversity. There are five metrics used to assess the time series components discussed in the previous section : Detrended Fluctuation Analysis (DFA), Spectral Entropy, Kurtosis, Skewness, and Gaussianity of the differenced values. They have been extensively used in the literature as individual assessment measures of time series data [10].

Evaluation Criteria

In this paper, the evaluation goal is to demonstrate the diversity of the generated time series, where the goal is to achieve maximum diversity and thus, we incorporate three main approaches to assess the degree of diversity. First, we use the histogram plot to visually observe the diversity of the generated time series for each feature on an individual basis. In a histogram plot, the x-axis represents the range of values for a given feature and the y-axis shows the number of time series that fall into each specific interval. Therefore, using the histogram plot, we can visually observe the distribution of the time series for each feature, individually. Note that we are not looking for a histogram with a uniform distribution of values but instead, attempt to determine the non-empty intervals, so that we obtain time series for *all* feature values.

Second, we use the multivariate entropy score to provide an accumulative diversity score across individual features to return a single score for diversity. While it represents the dominant approach in the literature, the problem with the multivariate entropy score is that it calculates diversity independent of inter-feature relationships. Therefore, we propose the third evaluation measure known as the *coverage rate* which provides a more reliable evaluation for diversity.

Visualizing Feature Metrics

The results presented in this section use the 53,637 generated time series as input. Each set of results presented in figures 5, 6, 7, 8 and 9 represents

an analysis (metric) of a specific feature across the entire dataset. For each feature, a histogram plot has been provided that visually illustrates the diversity of time series over the potential range of values for the corresponding feature. Using these histograms, we evaluate the diversity of each feature independent of each of the other features and this evaluation can be referred to as *feature specific* diversity. As the goal is to observe the least number of empty intervals, a perfectly diverse dataset is one that has no zero intervals over the range of the possible values for the given feature.
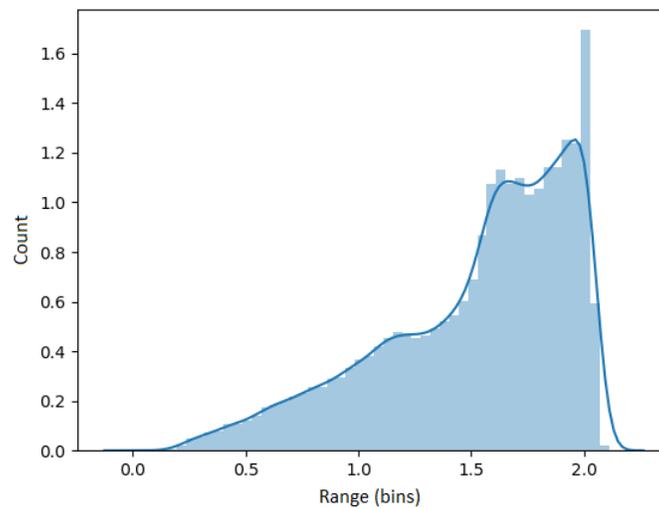
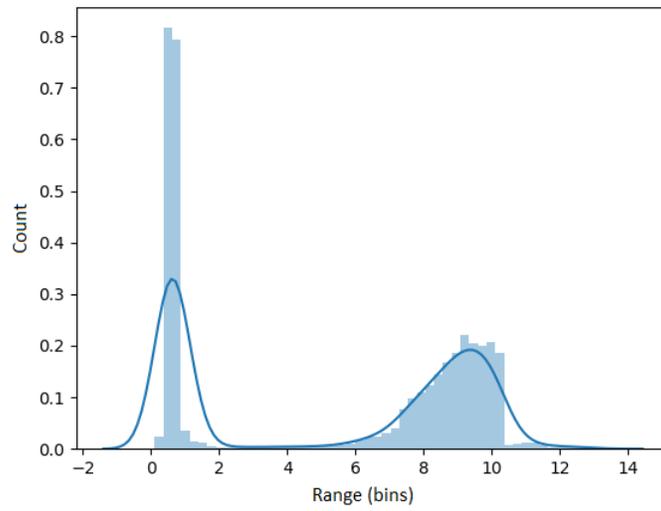Figure 5: Results for LRD Metric (DFA)
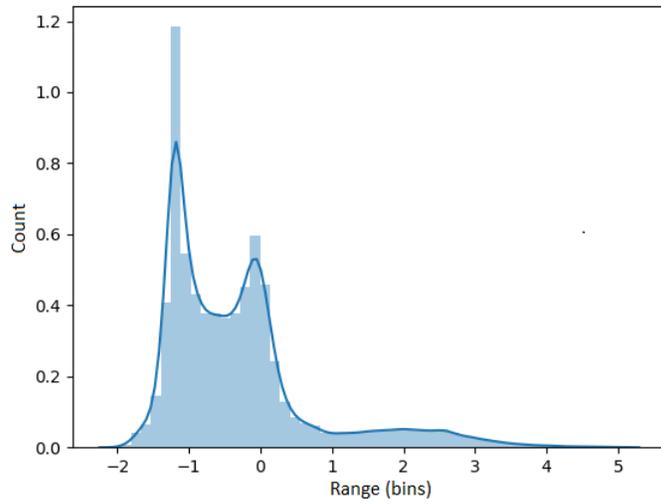
Figure 6: Results for Complexity

Figure 7: Results for Normality (Kurtosis)
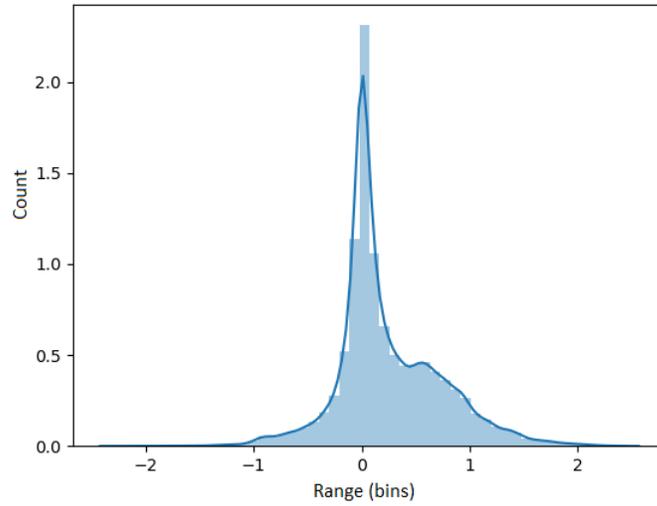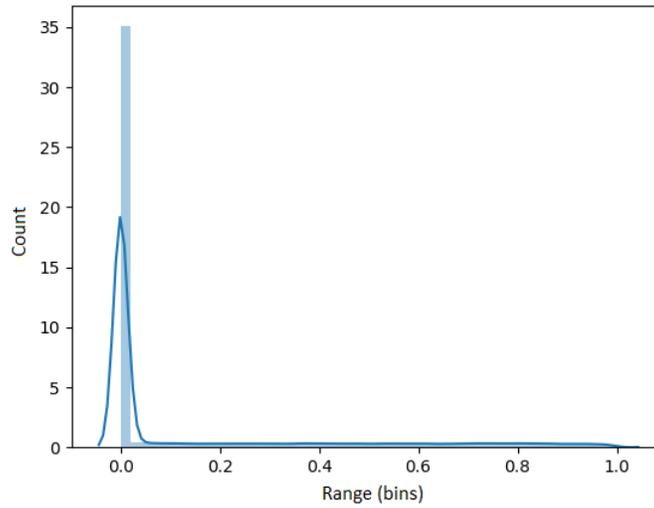
16

Figure 8: Results for Normality (Skewness)

Figure 9: Results for Normality (GoD)

Long range dependency for each series was calculated using a DFA analysis and shown in Figure 5. These results demonstrate that the synthetic series encapsulate all forms of long range dependency described by the DFA values in the previous section.

In Figure 6, the histogram for spectral entropy of the time series is presented. The entropy of values close to zero indicates high levels of self-similarity and thus, higher predictability. The results illustrate a high number of time series with low complexity (entropy close to zero) and also a high number of complex time series (entropy greater than 9). In addition, there are a lot of time series between these ranges, demonstrating that *all* complexity levels are present.

In Figure 7, the results for kurtosis show the expected diversity of negative (series of light tails or series of no outliers), zero (occasional outliers) and positive values (series of heavy tails or series with significant or numerous outliers). This is a strong indicator of diversity across the datasets.

The results for skewness are presented in figure 8, showing a high number of time series with negative skewness (series with a fatter or longer tails on the left side), zero skewness (series of symmetrical distribution) and positive skewness (series of heavy or long tails on the right side). Once again, this indicates a high level of diversity across the datasets.

Figure 9 illustrates the distribution of the gaussianity of the differences. A value of 1 indicates that the series follow a normal/Gaussian distribution and a value of 0 indicates no normality. The results show the generated series cover the entire range between zero and complete normality and thus, demonstrates a high level of diversity for the generated series.

Multivariate Entropy Score

The diversity measure in this paper is based on Shannon's entropy function which is frequently used to measure the amount of information in an encoded message [47], and shown in equation 9, where $x_1, x_2, ..., x_S$ are the possible values of X and $p(x_i)$ is the probability of observing x_i or $X = x_i$.

$$H(X) = -\sum_{i=1}^{S} p(x_i) \log p(x_i) \tag{9}$$

In order to measure diversity, we used a metric known as the *evenness* measure [47], which provides a normalized value for $H(X)$ based on its maximum, and shown in equation 10.

$$H_{max}(X) = -\sum_{i=1}^{S} \frac{1}{S} \log \frac{1}{S} = \log S \tag{10}$$

Therefore, the *diversity* of feature X is calculated by equation 11.

$$H_E(X) = \frac{H(X)}{H_{max}(X)} = -\frac{1}{\log S} \sum_{i=1}^{S} p(x_i) \log p(x_i) \qquad (11)$$

In this evaluation, we assume that all the features have equal significance, independent of the domain-specific constraints of the problem space. Assuming that all features have the same significance, the diversity for a multivariate (multi-feature) dataset with k features can be obtained using equation 12, where H will range between 0 and 1.

$$H = \frac{1}{k} \sum_{i=1}^{k} H_E(X^k) \qquad (12)$$

In order to implement this metric, each feature was categorized into buckets/zones as used traditionally by researchers. The categorization of the features, later shown in Table 3, is as follows:

- Spectral Entropy was categorized into three categories including A:$X < 1$, B:$1 \leq X < 9$ and C:$9 \leq X$.

- Kurtosis was categorized into three categories including A:$X < -0.3$, B:$-0.3 \leq X < 0.3$ and C:$0.3 \leq X$.

- Skewness was categorized into three categories including A:$X < -0.3$, B:$-0.3 \leq X < 0.3$ and C:$0.3 \leq X$.

- GoD was categorized into two categories including A:$X < 0.02$ and B:$0.02 \leq X$.

- DFA was categorized into seven categories including A:$X < 0.45$, B:$0.45 \leq X < 0.55$, C:$0.55 \leq X < 0.95$, D:$0.95 \leq X < 1.05$, E:$1.05 \leq X < 1.45$, F:$1.45 \leq X < 1.55$, G:$1.55 \leq X$.

Table 3 shows the breakdown of the proportion of series that belong to each of the categories outlined above. An N/A implies that this category is not appropriate for that metric.

The proportion of series that belong to each of the categories outlined above are shown in Figure 10

Table 3: Proportion of dataset relative to time series characteristics

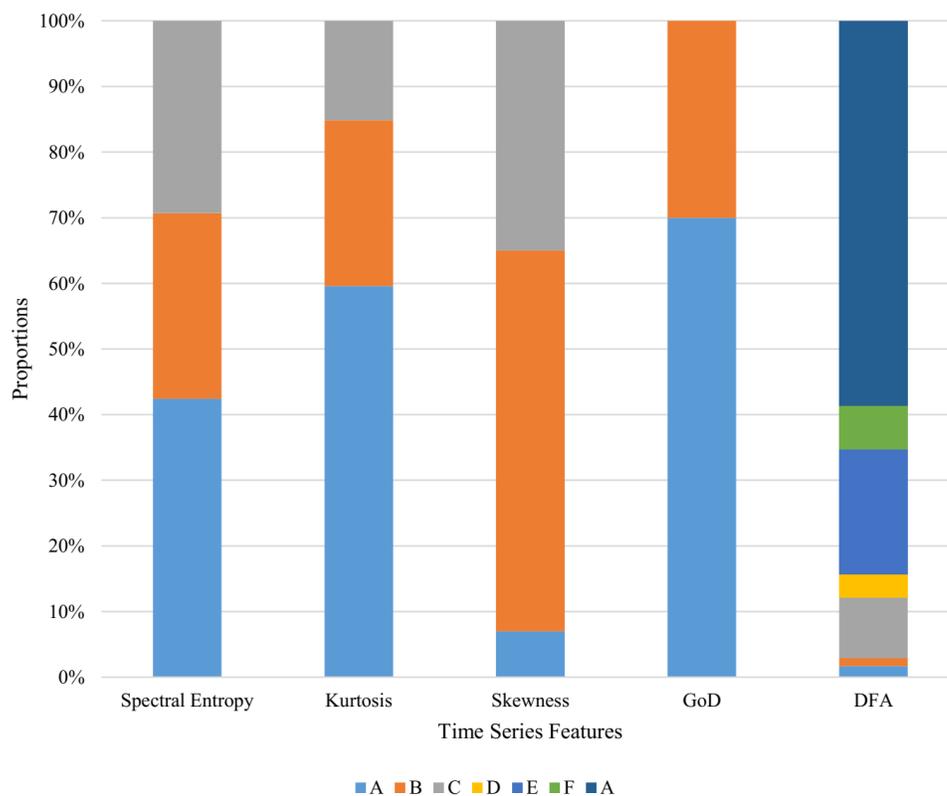| Feature	A	B	C	D	E	F	G
Spectral Entropy	0.42	0.28	0.29	N/A	N/A	N/A	N/A
Kurtosis	0.59	0.25	0.15	N/A	N/A	N/A	N/A
Skewness	0.07	0.58	0.35	N/A	N/A	N/A	N/A
GoD	0.70	0.3	N/A	N/A	N/A	N/A	N/A
DFA	0.017	0.012	0.092	0.035	0.190	0.067	0.584

Figure 10: Proportion of dataset relative to time series characteristics

Table 4 shows the H_{max} and H_E of each metric for the full dataset. These interim results are used to calculate the diversity as our final evaluation is to measure the diversity and coverage rate. The overall diversity score, H for

the dataset was 0.83. Here, $H(X)$, H_{max} and H_E which were obtained using equations 9, 10 and 11, and show that the level of diversity for each of the metrics examined ranges between 0.65 for DFA to 0.98 for Spectral Entropy. This is a significant result as it indicates the most diverse features or the features that have the best evenness. The low level of diversity for the DFA metric was predominantly due to the low levels of stationary data $DFA < 1$.

Table 4: H scores for each metric

Feature	$H(X)$	H_{max}	H_E
Spectral Entropy	1.55	1.58	0.98
Kurtosis	1.366	1.58	0.86
Skewness	1.25	1.58	0.79
GoD	0.88	1.00	0.88
DFA	1.837	2.80	0.65

The H score (diversity score) for each time series characteristics is shown in Figure 11
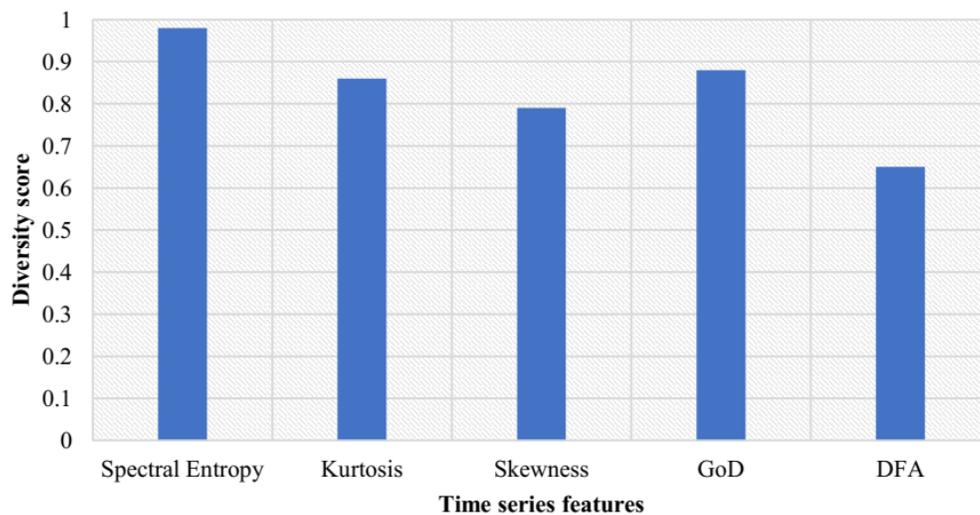

Figure 11: H scores for each metric

Feature Space Coverage

The feature space for the data is identified as all potential category combinations of the metrics outlined above. For this evaluation, we selected a measure of diversity that reflects the *percentage coverage* of the samples over the *potential* feature space. Using Table 3, there are: 3 categories for spectral entropy; 3 categories for Kurtosis; 3 categories for Skewness; 2 categories for GoD; and 7 categories for DFA. Thus, there is a total of $(3 \times 3 \times 3 \times 2 \times 7)$ 378 possible feature combinations, meaning 378 potential categories from our metrics. A full list of all feature combinations is provided in [18].

Figure 12 shows the number of time series where a specific category was represented by our synthetic data. Here, the *x-axis* represents all 378 possible categories, each using a unique *category_id*, and the *y-axis* shows the number of time series that fall under that *category_id*. For some categories, it is clear that there are multiple time series in the datasets whereas other categories are absent altogether.
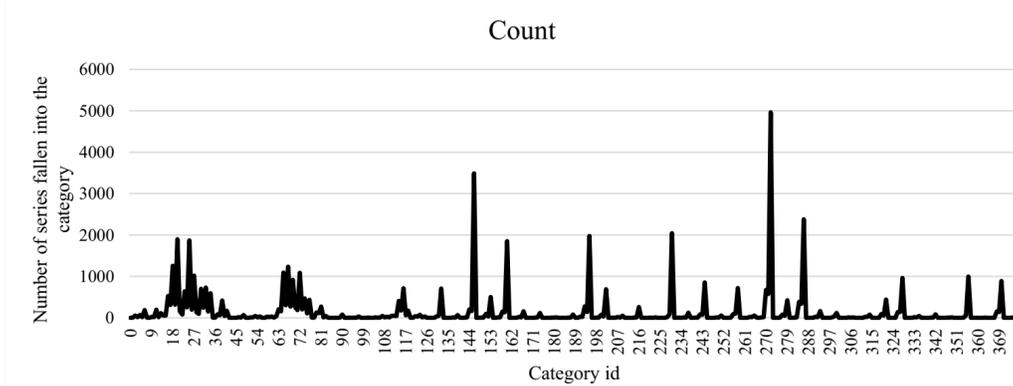

Figure 12: Number of series in each category

Conclusions

Researchers using time series data are often faced with the problem of insufficient data for the purposes of testing and validating their algorithms. In this work, we presented a methodology for the creation of a large number (53,637) of time series which are now available to the research community [18]. Their construction had an emphasis on *diversity* and a validation framework to ensure a robust evaluation of the synthetic time series created. Our method comprised 5 well-known time series features and used a multivariate

entropy measure to examine the diversity of the created time series based on these five features. The experimental results showed that our overall dataset measured diversity at 83.4%, which we believe to be a significant achievement. We have also proposed a new diversity assessment measure called the *coverage rate* which reflects the coverage of the dataset over the full feature space. The results show that our series exhibit a coverage rate of 72%, which delivers a significant contribution for such a large dataset.

There are some limitations to this research which we feel should be highlighted. Firstly, our paper considers only five features for studying diversity and future research could adopt more time series features into a more advanced study of diversity when building synthetic time series. Additionally, this research constructs only $50K$ time series and a wider set would be necessary to accommodate the additional features. Secondly, we assume that all features have equal significance, independent of the domain-specific constraints of the problem space. To advance our work, researchers could incorporate an additional customization step for determining the significance of features, applicable to *each* specific domain.

Acknowledgements

Declaration of the conflict of Interest

The authors declare that there are no known competing financial interests or personal relationships which could have appeared to affect the research presented in this paper.

References

[1] H. I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, P.-A. Muller, Deep learning for time series classification: a review, Data Mining and Knowledge Discovery 33 (2019) 917–963.

[2] Q. Zhang, J. Wu, H. Yang, Y. Tian, C. Zhang, Unsupervised feature learning from time series., in: IJCAI, New York, USA, 2016, pp. 2322–2328.

[3] H. Wang, Q. Zhang, J. Wu, S. Pan, Y. Chen, Time series feature learning with labeled and unlabeled data, Pattern Recognition 89 (2019) 55–66.

[4] F. Bahrpeyma, M. Roantree, A. McCarren, Multi-resolution forecast aggregation for time series in agri datasets (2017). URL: http://ceur-ws.org/Vol-2086/AICS2017_paper_24.pdf.

[5] H. Hayashi, T. Shibanoki, K. Shima, Y. Kurita, T. Tsuji, A recurrent probabilistic neural network with dimensionality reduction based on time-series discriminant component analysis, IEEE transactions on neural networks and learning systems 26 (2015) 3021–3033.

[6] S. B. Taieb, R. J. Hyndman, A gradient boosting approach to the kaggle load forecasting competition, International journal of forecasting 30 (2014) 382–394.

[7] A. Shamshad, M. Bawadi, W. W. Hussin, T. Majid, S. Sanusi, First and second order markov chain models for synthetic generation of wind speed time series, Energy 30 (2005) 693–708.

[8] M. Small, D. Yu, R. G. Harrison, Surrogate test for pseudoperiodic time series data, Physical Review Letters 87 (2001) 188101.

[9] M. Roantree, J. Liu, A heuristic approach to selecting views for materialization, Softw. Pract. Exp. 44 (2014) 1157–1179. URL: https://doi.org/10.1002/spe.2192. doi:10.1002/spe.2192.

[10] X. Wang, K. Smith-Miles, R. Hyndman, Rule induction for forecasting method selection: Meta-learning the characteristics of univariate time series, Neurocomputing 72 (2009) 2581–2594.

[11] G. E. Box, G. M. Jenkins, G. C. Reinsel, G. M. Ljung, Time series analysis: forecasting and control, John Wiley and Sons, 2015.

[12] Q. Zhang, J. Wu, P. Zhang, G. Long, C. Zhang, Salient subsequence learning for time series clustering, IEEE transactions on pattern analysis and machine intelligence 41 (2018) 2193–2207.

[13] M. Füllsack, M. Kapeller, S. Plakolb, G. Jäger, Training lstm-neural networks on early warning signals of declining cooperation in simulated repeated public good games, MethodsX (2020) 100920.

[14] A. Efstratiadis, Y. G. Dialynas, S. Kozanis, D. Koutsoyiannis, A multivariate stochastic model for the generation of synthetic time series at multiple time scales reproducing long-term persistence, Environmental Modelling & Software 62 (2014) 139–152.

[15] Y. Chen, Y. Wang, D. Kirschen, B. Zhang, Model-free renewable scenario generation using generative adversarial networks, IEEE Transactions on Power Systems 33 (2018) 3265–3275.

[16] A. M. Narasimhamurthy, L. I. Kuncheva, A framework for generating data to simulate changing environments., in: Artificial Intelligence and Applications, 2007, pp. 415–420.

[17] R. A. Ganjoei, H. Akbarifard, M. Mashinchi, S. A. M. J. Esfandabadi, A method for estimating width bands of variables in economics under uncertainty conditions, MethodsX (2020) 101184.

[18] F. Bahrpeyma, M. Roantree, P. Cappellari, M. Scriney, A. McCarren, Establishing diversity in synthetic time series for prediction performance evaluation, 2021. URL: `https://doi.org/10.5281/zenodo.4455631`. doi:10.5281/zenodo.4455631.

[19] M. Khashei, M. Bijari, G. A. R. Ardali, Improvement of auto-regressive integrated moving average models using fuzzy logic and artificial neural networks (anns), Neurocomputing 72 (2009) 956–967.

[20] T. Panagiotidis, Testing the assumption of linearity, Economics Bulletin 3 (2002) 1–9.

[21] Y.-W. Si, J. Yin, Obst-based segmentation approach to financial time series, Engineering Applications of Artificial Intelligence 26 (2013) 2581–2596.

[22] Y. Zhang, P. Shang, The complexity–entropy causality plane based on multivariate multiscale distribution entropy of traffic time series, Nonlinear Dynamics 95 (2019) 617–629.

[23] S. J. Taylor, Modelling financial time series, world scientific, 2008.

[24] H. D. Vinod, J. López-de Lacalle, et al., Maximum entropy bootstrap for time series: the meboot r package, Journal of Statistical Software 29 (2009) 1–19.

[25] A. Bagnall, A. Bostrom, J. Large, J. Lines, Simulated data experiments for time series classification part 1: accuracy comparison with default settings, arXiv preprint arXiv:1703.09480 (2017).

[26] L. Kegel, M. Hahmann, W. Lehner, Feature-based comparison and generation of time series, in: Proceedings of the 30th International Conference on Scientific and Statistical Database Management, 2018, pp. 1–12.

[27] Y. Kang, R. J. Hyndman, F. Li, Gratis: Generating time series with diverse and controllable characteristics, Statistical Analysis and Data Mining: The ASA Data Science Journal 13 (2020) 354–376.

[28] O. Mogren, C-rnn-gan: Continuous recurrent neural networks with adversarial training, arXiv preprint arXiv:1611.09904 (2016).

[29] H.-W. Dong, W.-Y. Hsiao, L.-C. Yang, Y.-H. Yang, Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[30] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, D. N. Metaxas, Stackgan++: Realistic image synthesis with stacked generative adversarial networks, IEEE transactions on pattern analysis and machine intelligence 41 (2018) 1947–1962.

[31] M. Alzantot, S. Chakraborty, M. Srivastava, Sensegen: A deep learning architecture for synthetic sensor data generation, in: 2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), IEEE, 2017, pp. 188–193.

[32] P. J. Brockwell, R. A. Davis, Introduction to time series and forecasting, springer, 2016.

[33] F. Bahrpeyma, Multistep ahead time series prediction, Ph.D. thesis, Dublin City University, 2021.

[34] P. Doukhan, G. Oppenheim, M. Taqqu, Theory and applications of long-range dependence, Springer Science & Business Media, 2002.

[35] M. Theodosiou, Disaggregation & aggregation of time series components: A hybrid forecasting approach using generalized regression neural networks and the theta method, Neurocomputing 74 (2011) 896–905.

[36] M. Fernández-Martínez, J. L. G. Guirao, M. Á. Sánchez-Granero, J. E. T. Segovia, Fractal Dimension for Fractal Structures: With Applications to Finance, volume 19, Springer, 2019.

[37] M. Tarnopolski, On the relationship between the hurst exponent, the ratio of the mean square successive difference to the variance, and the number of turning points, Physica A: Statistical Mechanics and its Applications 461 (2016) 662–673.

[38] J. Beran, Statistics for long-memory processes, Routledge, 2017.

[39] G. P. Zhang, M. Qi, Neural network forecasting for seasonal and trend time series, European journal of operational research 160 (2005) 501–514.

[40] R. Hyndman, The interaction between trend and seasonality, International Journal of Forecasting 20 (2004) 561–563.

[41] H. E. Hurst, Long term storage capacity of reservoirs, ASCE Transactions 116 (1951) 770–808.

[42] S. Hergarten, Self organized criticality in earth systems, Springer, 2002.

[43] J. Sethna, Statistical mechanics: entropy, order parameters, and complexity, volume 14, Oxford University Press, 2006.

[44] J. S. Richman, D. E. Lake, J. R. Moorman, Sample entropy, in: Methods in enzymology, volume 384, Elsevier, 2004, pp. 172–184.

[45] D. Ruppert, Statistics and finance: An introduction, Springer, 2014.

[46] F. Bahrpeyma, M. Roantree, A. McCarren, Multistep-ahead prediction: A comparison of analytical and algorithmic approaches, in: International Conference on Big Data Analytics and Knowledge Discovery, Springer, 2018, pp. 345–354.

[47] T. Pham, R. Hess, C. Ju, E. Zhang, R. Metoyer, Visualization of diversity in large multivariate data sets, IEEE Transactions on Visualization and Computer Graphics 16 (2010) 1053–1062.