AIMS Mathematics, 8(10): 24528–24544.
DOI: 10.3934/math.20231251
Received: 20 June 2023
Revised: 21 July 2023
Accepted: 31 July 2023
Published: 18 August 2023

http://www.aimspress.com/journal/Math

Research article

Data augmentation in economic time series: Behavior and improvements in predictions

Ana Lazcano de Rojas*

Universidad Francisco de Vitoria, Faculty of Law and Business, Spain

* **Correspondence:** Email: ana.lazcano@ufv.es.

Abstract: The performance of neural networks and statistical models in time series prediction is conditioned by the amount of data available. The lack of observations is one of the main factors influencing the representativeness of the underlying patterns and trends. Using data augmentation techniques based on classical statistical techniques and neural networks, it is possible to generate additional observations and improve the accuracy of the predictions. The particular characteristics of economic time series make it necessary that data augmentation techniques do not significantly influence these characteristics, this fact would alter the quality of the details in the study. This paper analyzes the performance obtained by two data augmentation techniques applied to a time series and finally processed by an ARIMA model and a neural network model to make predictions. The results show a significant improvement in the predictions by the time series augmented by traditional interpolation techniques, obtaining a better fit and correlation with the original series.

Keywords: time series forecasting; financial forecasting; data augmentation
Mathematics Subject Classification: 68T07;68T09

1. Introduction

The performance of neural networks for time series prediction is conditioned by the amount of data available to train the network. The amount of data available plays a crucial role in the accuracy and reliability of the predictions made by statistical models. This aspect acquires significant importance in the academic and business fields, where the ability to correctly predict future patterns and behaviors from a time series is essential for strategic decision making.

An insufficient number of observations can lead to unrepresentativeness of the underlying patterns and trends in the data. Time series are typically made up of data that evolves over time and a limited number of observations may not adequately capture the inherent variability and complexity of this data. These aspects are fundamental for the construction of robust and accurate prediction models, since they provide crucial information about the behavior of the data. The literature has highlighted the importance of having an adequate number of observations to obtain reliable results in the prediction. For example, Box et al. [1] emphasized the need for at least 30 observations to adequately model a time series and make accurate forecasts. Similarly, Shumway and Stoffer [2] suggested that a minimum of 50 to 100 observations are necessary to build reliable models and avoid wrong conclusions. The development of data augmentation techniques allows solving this problem, providing the network with enough training data and consequently more accurate predictions. These data augmentation methods employ different techniques to augment time series observations, using estimates and predictions that generate data similar to the original. Economic time series are characterized by having a trend, autocorrelation, seasonality and being mostly non-stationary, which highlights the importance of achieving adequate data augmentation techniques that allow generating new observations without affecting the characteristics of the series.

In their study Iwana and Uchida [3] conducted a review of the literature regarding the use of data augmentation algorithms for time series classification. In their research, they reviewed 12 methods to achieve data improvement and 128 sets of classification and evaluation with 6 types of neural networks. However, the most recent literature focuses on the use of other types of techniques such as the use of GAN networks. Therefore, this study lacks a comparison with traditional data augmentation approaches.

The work carried out by Iglesias et al. [4] tries to compile various data augmentation techniques oriented to time series, allowing an exhaustive comparison of the methodologies in the domain of time series.

Research by Liu et al. [5] propose various methodologies aimed at increasing time series data such as AddNoise, permutation, scaling and warping, verifying these methods through two deep learning models with real-time time series, showing an improvement in classification tasks. From these investigations arises the need to make a comparison between the different data augmentation techniques and provide information on the performance of the resulting time series for prediction tasks.

The main objective of this research is to identify the most suitable model to increase the number of observations of economic time series, which allows to generate a consistent number of observations, allowing to improve the performance of classical models and artificial neural networks (ANN) to make predictions in financial time series. After carrying out multiple experiments, it is observed how interpolation techniques achieve a substantial improvement in the error metrics of different prediction models, allowing more reliable forecasts to be generated when sufficient data is lacking.

2. Related work

2.1. Time series augmentation

In recent decades, data augmentation algorithms have become highly relevant in changing machine learning and artificial intelligence. These algorithms allow the generalization and improvement of the performance of the trained models. They are used to generate new synthetic data

from existing ones through the application of simple transformations or the generation of synthetic data through complex generative models.

Various authors deal with the relevance of the number of observations necessary to obtain good predictions, carrying out experiments in which the performance of the models is measured based on this number and how the error metrics vary ([6–8]) finding significant differences between the observations they consider necessary, varying between 100 and 300.

The use of data augmentation algorithms dates back to the 1990s, when researchers began using these techniques with the goal of increasing and improving the quality and quantity of data available to machine learning models. One of the first studies in the area was carried out by Lecun et al. [9], in which they presented a method that allowed the generation of images from existing ones, allowing a significant improvement in the performance of image classification models.

Over the next decade, more sophisticated data augmentation methods, such as slice [10] and rotation [11], were developed for augmenting image data sets using the application of simple transformations.

Data augmentation algorithms are also used in the generation of texts. Data augmentation techniques based on the substitution of synonyms and the random elimination of words have been developed in [12].

In this study, Wong and Leung [13] carried out an exhaustive review of the existing data augmentation techniques for neural networks. In the research they describe different data augmentation techniques such as rotation, translation and scaling, among others. Subsequently, the research carried out in [14] explored the deformation-based data augmentation technique by applying nonlinear deformations in small windows of the time series to generate new series with different patterns of variation.

This data augmentation technique is introduced in the field of time series, allowing to increase the size of these and improve the performance of predictions by neural networks. Authors like Yoon et al. [15] present a data augmentation technique based on generative adversarial networks (GAN) for use in time series. The TimeGAN model is trained with the objective of generating new time series from an existing data set and it is used to generate new time series that allow preserving the statistical characteristics of the original data set.

Different data augmentation techniques for time series are investigated by researchers in [16,17], where they propose different techniques that allow increasing the number of observations of the time series through generative models, allowing to improve the accuracy of the predictions through neural network models.

2.2. Data augmentation for time series

Throughout the last few years, different algorithms have been developed that allowed the increase in the time series data available to perform the training of the neural networks. This increase allows minimizing the risk of overfitting by the network in case of lack of sufficient data and improve the accuracy of the models. There are several techniques to achieve the increase in available observations in a time series.

Most recent research is focused on augmenting data from images, video or natural language processing (NLP) ([18–20]). These techniques are focused on correcting the imbalance or lack of data in the information sets, but there are other application areas where these problems are more common, such as time series and especially economic ones.

The growing interest in this type of technique has allowed innovation in methodologies. Data imputation techniques make it possible to increase the size of time series using various techniques such as interpolation, extrapolation, smoothing, the use of regression techniques and the use of machine learning models to predict missing values. These include extrapolation, resampling, transformations, interpolation, imputation and simulation. These techniques are reviewed in research such as those carried out in [21], where they review data augmentation techniques, including imputation techniques. García-Molina et al. [22] carried out a study in which, using imputation techniques, they completed the missing values in a time series, managing to increase the size of the data set.

Interpolation is a technique by which new data is generated from existing data. This technique consists of generating new data points within the range of existing values by estimating values from known data. Guennec et al. [23] proposed the use of this to increase the observations of a time series for the training of a convolutional neural network for the classification of time series. The interpolation technique is commonly used in time series modeling to achieve an increase in the available observations, some of the best known methods are spline and Fourier.

Through the extrapolation technique it is possible to achieve the generation of new data, in this case outside the range of existing values, the Holt-Winters and Box-Jenkins techniques are the best known. In their research, Salinas et al. [24] compared this model with other classic and deep learning models. Gashler and Ashmore [25] explored the power of these models for data augmentation.

One of the simplest forms of extrapolation is linear extrapolation, which uses a straight line to extend the trend of the observed data. This model can be represented as:

$$y = mx + b \tag{1}$$

where y represents the value to be predicted, x represents the independent variable corresponding to the value to be predicted, m is the slope of the straight line and b is the y-intercept.

Monte Carlo simulation will perform data augmentation from a probability distribution. It involves the generation of a large number of random values from a given probability distribution and will be used to generate new data. Bootstrap and Marcoval are the most used techniques in Monte Carlo simulation. Studies like presented in [26,27] based their research on the use of this technique to achieve an increase in the number of observations of a time series using machine learning techniques.

The Monte Carlo simulation is based on the fact that given a data set of size n: $X = \{x1, x2, ..., xn\}$ the probability distribution of the data $p(X)$ will be estimated, a set of m random numbers from the estimated probability distribution $q(X)$ and adjusting the generated values to match the range and scale of the original data.

Then, the likelihood of getting the sample from the distribution is given by the Eq (2):

$$L(\theta) = f\theta(x1,x2,\ldots,xn|\theta) \tag{2}$$

Let θ be the parameter vector for f, which can be either a probability mass function (PMF) for discrete distributions or a probability density function (PDF) for continuous distributions. We will denote the pdf/pmf as $f\theta$. Let the sample drawn from the distribution be $x1,x2,...,xn$.

Another of the most used techniques to achieve an increase in the number of observations in a time series is the increase in synthetic data, which through random transformations to existing data will achieve the generation of new data. These transformations can include rotations, translations, scaling and deformations, often using the wavelet transform. Investigations such as those carried out in [28] and [29] show the usefulness of this type of model, combined with the use of machine learning

techniques to generate new data.

The generation of synthetic data through deep learning (DL) techniques has proven to be highly effective, being able to distinguish between different techniques used for this purpose:

Variational autoencoder (VAE): An autoencoder neural network will learn to encode and decode data. While the encoder maps the input data to a distribution of latent variables, the decoder will map the latent variables to the desired output in order to maximize the probability of the original data given the parameters of the latent distribution and minimize the difference between the two. This function will allow you to generate new data similar to the originals, including slight modifications.

This technique was developed in [30]. They presented an architecture that combined the use of an autoencoder with a Bayesian inference probabilistic model. The research by Deng [31] presented a variant that allowed learning multimodal latent representations of the data.

An encoder is capable of compressing an input x into a latent space z, while a decoder will act from that latent space z to obtain the reconstruction, it can be defined as:

$$p(x) = \sum_z p(x|z)p(z) \tag{3}$$

In the formula above, x is training data and z is the hidden feature that cannot be observed in x data.

Generative adversarial networks (GAN): These neural networks proposed by Goodfellow et al. [32] are trained to generate new data from a random noise distribution. Through two networks, a generator and a discriminated one, the first will generate the synthetic data while the second will classify whether the data is real or synthetic ([15]). The goal is for the generated data to be realistic enough that the discriminator is unable to distinguish it from the actual data. Research by Isola et al. [33] proposed a variant called conditional GAN, which allows controlling the output of the generator through a conditional input. The formula for the entire GAN is as follows:

$$\min_G \max_D V(D,G) = E_{X\sim p_{data}(x)}[\log(D(x))] + E_{z\sim p_z(z)}\left[\log\left(1 - D\left(G(z)\right)\right)\right] \tag{4}$$

where E is the expected value of the distribution function, $p_{data}(x)$ is the distribution of the real sample and $p\_((z))$ is the distribution that generates the sample [34].

The generation of synthetic data through recurrent neural networks (RNN) and long short-term memory (LSTM) developed by Hochreiter and Schmidhuber [35], will allow the generation of new data by predicting the next element in the sequence from the previous elements.

WaveNet consists of a convolutional neural network (CNN) will work similar to RNN and LSTM but through an autoregressive convolutional network. In the research by Oord et al. [36] used this technique to generate high-quality signals in an autoregressive way.

Other generative modeling techniques for time series are generative flow, which was proposed in [37] in 2014, and transformers, which have been used for text generation and adapted for time series in [38] in 2019.

2.3. Data augmentation for time series

Mathematical methods for time series forecasting have constantly evolved since the beginning of research in this field. Poynting [39] tried in his study to eliminate the trend and cyclical fluctuations by averaging over a given time interval. Later, other researchers wrote in [40,41] about the elimination of trends by including high-order polynomials.

In recent years, the development of time series forecasting techniques has been developed and applied in a wide variety of fields, from finance and economics to engineering and physics. In its beginnings, prediction techniques were based on classical statistical models such as exponential smoothing models [42] or ARIMA models developed in [1], who through adjustment and smoothing techniques make the predictions of the series based on their historical values.

In the 1980s, artificial neural network (ANN) models were included among the prediction techniques, which have the ability to model non-linearly complex relationships between the input and output variables, which makes them useful for this type of prediction [43].

The amount of data currently available makes neural networks a dominant technique for which there is an extensive literature. Zhang et al. [44] carry out an extensive review of the investigations in which neural network techniques have been used for the prediction of time series, concluding with the superiority of neural networks over classical techniques and this is due to the properties that characterize the networks. artificial neurons.

One of these features, described in [45] is the ability to generalize data and make predictions from it.

One of the first researchers to apply ANN to time series forecasting was presented in [46], in his work he predicted gold prices using backpropagation neural networks. Subsequently, different types of ANN have been developed, among which are recurrent neural networks (RNN), especially long short term memory (LSTM) networks developed by Hochreiter and Schmidhuber [35], which allow modeling long-term temporal dependencies.

In the same line of research, in [47] was studied the behavior of artificial neural networks for the prediction of time series comparing with six statistical methods, resulting in a better result of artificial neural networks and achieving more accurate predictions.

This evolution has an impact on an improvement in the predictions, Siami-Namini and Namin [48] compared the performance of an ARIMA model with that of an LSTM model, finding lower error metrics on the part of the ANN model for the prediction of industrial production in Taiwan.

In recent decades, research has focused on hybrid models that allow combining the advantages of traditional statistical models with ANN, achieving greater precision in the results.

The research carried out by Ravi et al. [49] focused on a hybrid model that combined the ARIMA technique with the use of neural networks for the prediction of inflation in the United States (USA), demonstrating superior performance by of the hybrid model with respect to the ARIMA and ANN models individually.

In this line, various authors have developed their research around the combination of different types of neural networks with the aim of improving these predictions ([50–54]).

3. Experiments

3.1. Data description

In this section the data used in the experiment will be described.

To observe the behavior of the economic time series in the face of an increase in data carried out with different techniques, experiments are carried out starting from three different data sets. These time series vary in the number of source observations as well as in the trend and characteristics of the data, allowing a global vision of how the increase in data influences this type of data.

In first place, the time series corresponding to the Morgan Stanley capital international (MSCI)

index obtained from Thomson Eikon Reuters was obtained with a daily timeframe with data from November 15, 2007 to August 12, 2022.

The MSCI index is used as a reference to evaluate the performance of equity markets worldwide and is made up of a series of indices that cover different geographical regions and sectors of the economy. The index is considered a key indicator of stock market performance and is closely watched by investors and financial analysts.

The second data set used is the China containerized freight index (CCFI). It is a measure used in the field of container shipping to assess changes in ocean freight prices on trade routes that involve China. Specifically, the CCFI focuses on freight prices for export containers from Chinese ports to destinations around the world.

The CCFI index has become a highly relevant tool for the shipping industry and for stakeholders involved in international trade, as it provides key information on trends and fluctuations in ocean freight costs. Allowing companies and industry analysts to make informed decisions regarding logistics planning, cost management and competitiveness assessment.

This time series consists of 2855 observations between August 2, 2011 and August 11, 2022 with a daily temporality.

Lastly, the data augmentation models have been tested with the time series corresponding to the gold price index, through the Gold Spot Price index, it is a reference used to determine the current value of gold in the spot market. Refers to the price at which gold can be bought or sold immediately, with immediate delivery and cash settlement. This index is widely followed and used as a key indicator of the price of gold in real time.

The spot price of gold is influenced by various factors, including supply and demand, global economic conditions, monetary policy, geopolitical stability and financial market movements. The gold supply comes from mining production, as well as the sale of recycled gold and from central banks.

For this time series, the available data is from May 13, 2015 to June 15, 2022, with 1850 observations.

3.2. Dataset augmentation

To carry out this experiment, two different data augmentation techniques are used with the aim of verifying the performance in the prediction with neural networks of both techniques with an economical time series.

3.2.1. Interpolation

First, the interpolation technique is chosen. The interpolation technique is based on using mathematical methods to estimate unknown or missing values in a time series, based on the values observed at earlier and later moments. It can be used to fill in gaps in a time series, that is, to add missing observations, thus increasing the amount of data available for analysis and prediction.

The main objective of this technique is to seek to find a function that allows to approximate the missing values in a coherent way with the observed values. The interpolation method can be defined as:

$$y_i = \sum_{j=i-k}^{i+k} \frac{(j-i+d_i)(i+k-d_i-j)}{2k^2} y_j \tag{5}$$

where k represents the number of observed values on each side of i and di is the distance between i and the closest observed value. This technique calculates the missing value as a linear combination of the closest observed values, weighted by a quadratic function that favors values that are closer together and assigns less importance to values that are further away. Figure 1 reflects how the interpolation technique adds values in intermediate positions of the time series, generating a series similar to the starting one, but significantly increasing the number of observations. The three time series are different from each other, the shocks and the main characteristics of the series remain, but the observations for the same date change significantly.

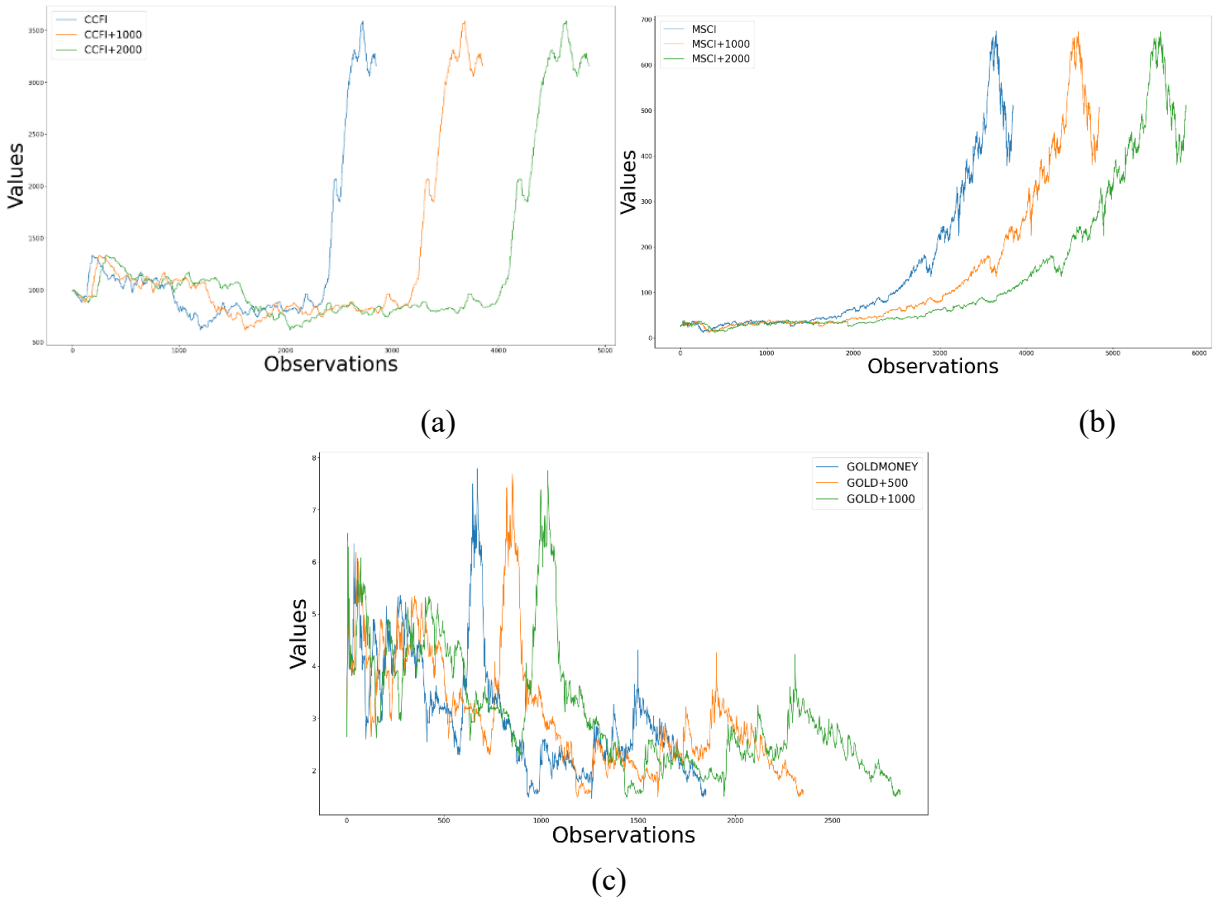


Figure 1. Interpolation augmentation of the time series. (a) CCFI;(b) MSCI; (c) GOLD.

3.2.2. Interpolation

The Tsaug library for time series data augmentation, implemented in Python, provides a wide variety of transformations applicable to the data in order to generate new observations, some of the transformations include time shift, amplitude shift, noise reduction, smoothing and frequency shift.

This implementation is based on the creation of transformation objects, applied to the input time series, applying a specific operation to a time series.

Tsaug also provides composition functions to combine transformations sequentially or randomly. This allows you to create complex sequences of transformations.

Some studies have shown that data augmentation using TSAUG can significantly improve the

performance of neural network models in time series forecasting. For example, Wang et al. [55] used TSAUG to augment the electric power time series dataset and achieved a significant improvement in the performance of neural network models. Several authors have demonstrated its usefulness in augmenting time series for classification tasks ([23,56,57]). TSAUG has also been shown to be effective in predicting financial time series [58] and predicting medical time series [59].

Figure 2 shows how the TSAUG technique adds the observations to the end of the time series, which allows working with the same observations for training, while in the test the time series will use different values when incorporating the new observations in the final part of the series. This technique will perform shock replay over the newly incorporated data.

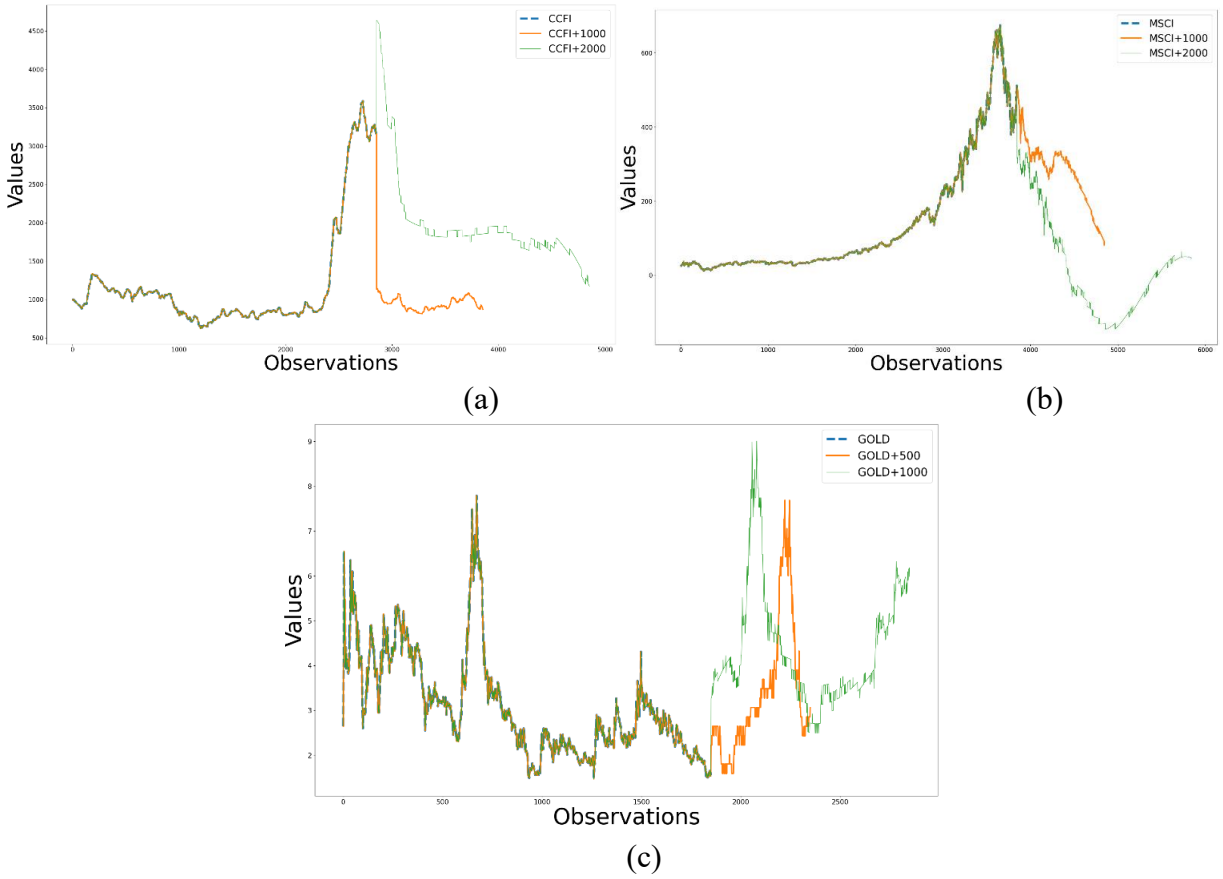


Figure 2. TSAUG augmentation of the time series. (a)CCFI; (b) MSCI; (c) GOLD.

3.3. Evaluation metrics

For the evaluation of the performance of the data augmentation presented in this work, the four most used error metrics have been chosen:

Root mean squared error (RMSE):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_t - \hat{Y}_t)} \tag{6}$$

Mean squared error (MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_t - \hat{Y}_t)^2 \tag{7}$$

R-squared (R2):

$$R^2 = 1 - \frac{\sum_{i=1}(Y_t - \hat{Y}_t)^2}{\sum_{i=1}(Y_t - \hat{Y})^2} \quad (8)$$

RMSE is a metric used to measure prediction error, or neural network performance. The lower this value, the better the prediction.

The value of R2 is closely related to his MSE and is also used to assess model performance. The output independent variable predicts the amount of variation in the output dependent attribute. The closer this value is to 1, the better the network performance.

The use of RMSE and MSE is justified due to their ability to measure the average difference between the predictions and the actual values. These metrics quantify the mean square error and the square root of the square error, respectively. Both metrics penalize the largest errors and provide a measure of the model's accuracy in terms of the spread of the errors [60].

These metrics have been widely adopted in the scientific literature and have been used in various research fields, such as stock price forecasting and electricity demand forecasting [61].

3.4. Experimental results and discussion

3.4.1. *Data augmentation results*

The result of the increase in the data of the different time series is reflected in Table 1. For each time series, two tests are carried out. In the first, the data set is increased by 1000 observations for the CCFI and MSCI time series (26% and 35% respectively) and 500 for the GOLD series, 27%. In the second test, there is an increase of 2,000 observations for MSCI and CCFI and 1,000, being 52% and 70% for each series. For the GOLD series, there is an increase of 1,000, which represents 54% of the total.

Table 1. Data augmentation results.

Data	**Original**	**1º**	**2º**
MSCI	3846	4846	5846
GOLD	1850	2350	2850
CCFI	2855	3854	4854

3.4.2. Forecasting model results

The increase in data is carried out using the interpolation and TSAUG techniques. Later, the results are analyzed after processing all the time series with the ARIMA models and a multilayer perceptron network. Then, it will be possible to evaluate how the increase in observations of a time series makes it possible to improve the predictions produced by the models and the techniques that best model this increase.

Tables 2 and 3 reflect the results obtained by the ARIMA model after the different data increases carried out, reflecting an improvement in the metrics with the interpolation technique.

Table 2. TSAUG augmentation with ARIMA model.

Data	ORIGINAL			1º			2º		
	RMSE	MSE	R2	RMSE	MSE	R2	RMSE	MSE	R2
MSCI	**5.951**	**35.419**	**0.9988**	7.871	61.955	0.998	6.639	44.071	0.998
GOLD	**0.089**	**0.008**	**0.9598**	0.124	0.015	0.986	0.126	0.016	0.991
CCFI	**15.403**	**237.246**	**0.9996**	38.344	1470.250	0.997	27.641	764.028	0.998

Table 3. Interpolation augmentation with ARIMA model.

Data	ORIGINAL			1º			2º		
	RMSE	MSE	R2	RMSE	MSE	R2	RMSE	MSE	R2
MSCI	5.951	35.419	0.9988	3.571	12.750	0.9995	**2.663**	**7.091**	**0.9997**
GOLD	0.089	0.008	0.9598	0.066	0.004	0.9937	**0.058**	**0.003**	**0.9969**
CCFI	15.403	237.246	0.9996	12.153	147.693	0.9997	**9.737**	**94.801**	**0.9998**

It is observable how the increase in the ARIMA model is capable of better modeling the time series increased with the interpolation technique and that with an increase in the amount of data up to figures greater than 50%, an improvement in precision of 44% is achieved. In the RMSE metric for the MSCI time series, 65% for the GOLD series and 63% for the CCFI series. When increasing the data with the TSAUG library the metrics worsen substantially, especially in the first round of increasing observations.

Tables 4 and 5 show the results obtained with a multilayer perceptron model after increasing the observations in all time series.

Table 4. TSAUG augmentation with MLP model.

Data	ORIGINAL			1º			2º		
	RMSE	MSE	R2	RMSE	MSE	R2	RMSE	MSE	R2
MSCI	**3.797**	**14.416**	**0.9990**	3.931	15.451	0.9986	7.223	52.169	0.9899
GOLD	**0.033**	**0.001**	**0.9963**	0.054	0.003	0.9986	0.048	0.002	0.9976
CCFI	**11.338**	**128.539**	**0.9998**	1.517	2.300	0.9995	0.504	0.254	0.9999

Table 5. Interpolation augmentation with MLP model.

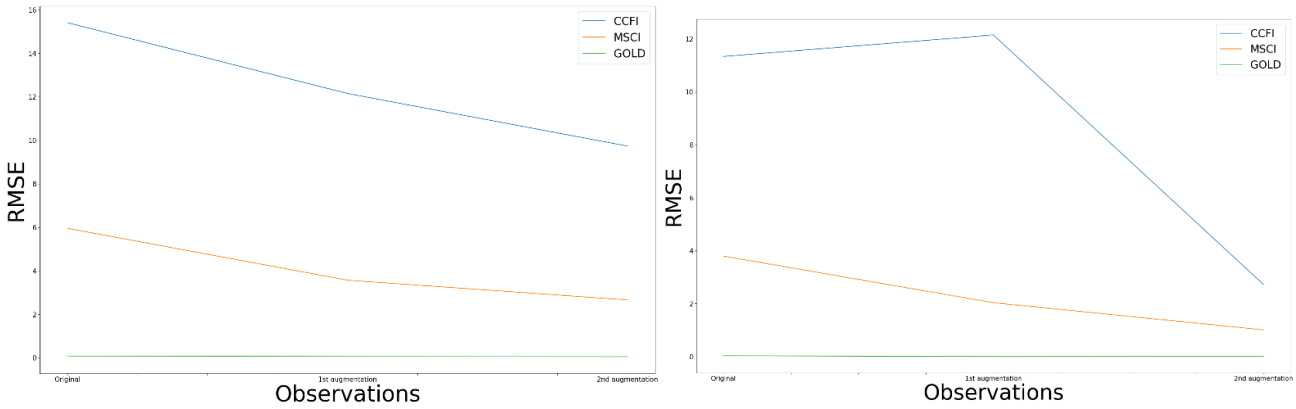
	ORIGINAL			1º			2º		
	RMSE	MSE	R2	RMSE	MSE	R2	RMSE	MSE	R2
MSCI	3.797	14.416	0.9990	2.035	4.142	0.9997	**1.010**	**1.020**	**0.9999**
GOLD	0.033	0.001	0.9963	0.002	0.043	0.9938	**0.003**	**0.000**	**0.9999**
CCFI	11.338	128.539	0.9998	12.146	147.520	0.9998	**2.721**	**7.405**	**0.9999**

As with the ARIMA model, the MLP artificial neural network is able to better model time series augmented with the interpolation technique. In this case the improvements in the second round of increases for the RMSE metric are 73%, 91% and 76% respectively.

Figure 3 shows the results observed in Tables 3 and 5, a decreasing trend as a function of the increase in the number of observations. When this increase occurs, the RMSE metric decreases

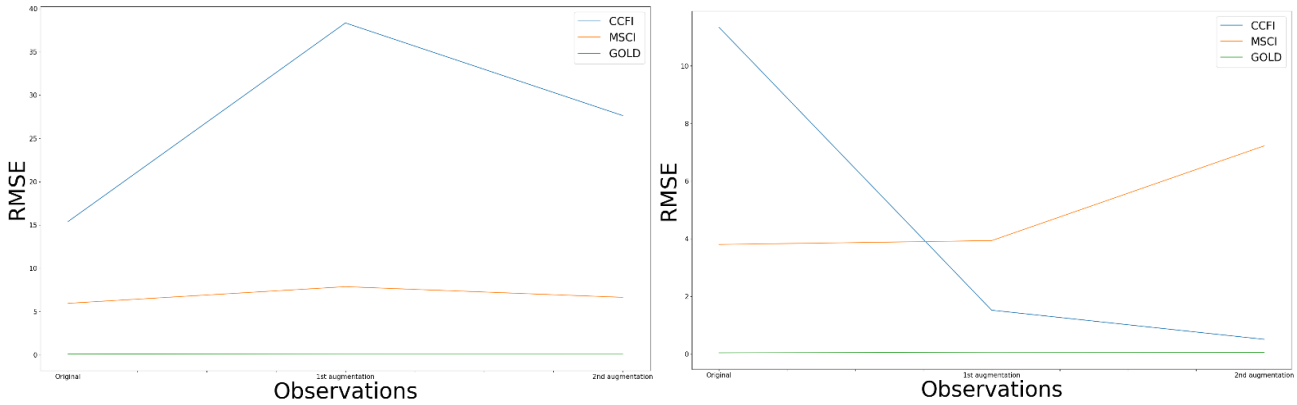
significantly, providing more accurate predictions and allowing model improvement.

On the other hand, Figure 4 shows how a greater number of observations does not directly imply a lower error, resulting in a less efficient technique for economic time series.



(a) (b)

Figure 3. Interpolation results using (a) ARIMA model; (b) MLP.



(a) (b)

Figure 4. TSAUG results using (a) ARIMA model; (b) MLP.

4. Conclusions and future lines of research

This research aims to analyze the impact of data augmentation techniques on economic time series and how it affects prediction algorithms, filling the gap in the existing literature that compares different data augmentation algorithms on economic time series.

The process of increasing data in time series is subject to the need to maintain the initial characteristics of the series so as not to interfere with the performance of predictions and statistical studies, achieving stability between the original time series and the increased time series. This study compares the behavior of different data augmentation algorithms applied to economic time series and the results of making predictions with classical statistical techniques and traditional techniques.

Two data augmentation techniques are used to increase the observations of three economic time series with the purpose of observing if this increase has an impact on a better fit of the prediction models.

Subsequently, they are analyzed by an ARIMA model and a multilayer perceptron (MLP) ANN model, to observe how the use of augmented economic time series affects the performance of the models.

The results obtained reflect how the imputation models allow an increase in observations, subsequently improving the metrics returned by both the ARIMA and MLP models The observations generated by the TSAUG library result in worse modeling, causing the metrics to show a higher error and not improving the training and testing process in either of the two cases.

These results coincide with the existing literature, highlighting the importance of increasing the amount of data available, presenting various data augmentation strategies used to improve the accuracy of forecast models [62].

Similarly, the conclusions coincide with Asem et al. [63], who through several experiments using the interpolation technique evaluated the effectiveness of these techniques in the generation of new data points to improve the accuracy of time series forecast models. The use of data augmentation techniques implies an improvement in the performance of the prediction techniques, improving the error metrics and therefore the predictions; this improvement is especially significant by using the interpolation technique. Based on the research carried out in this work, future lines are proposed in which combined approaches to increase data are sought, allowing an improvement in the performance of prediction models.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Conflict of interest

All authors declare no conflicts of interest that could affect the publication of this paper

References

1. G. E. Box, G. M. Jenkins, G. C. Reinsel, *Time series analysis: Forecasting and control*, Holden-Day, 1970.
2. R. H. Shumway, D. S. Stoffer, *Time series analysis and its applications: with R examples*. Springer, 2017. https://doi.org/10.1007/978-3-319-52452-8_3
3. B. K. Iwana, S. Uchida, An empirical survey of data augmentation for time series classification with neural networks, *PLoS ONE* **16** (2021), 0254841. https://doi.org/10.1371/journal.pone.0254841
4. G. Iglesias, E. Talavera, Á. González-Prieto, A. Mozo, S. Gómez-Canaval, Data Augmentation techniques in time series domain: a survey and taxonomy, *Neural Comput. Appl.*, **35** (2023), 10123–10145. https://doi.org/10.1007/s00521-023-08459-3
5. B. Liu, Z. Zhang, R. Cui, Efficient time series augmentation methods, In: 2020 13th international congress on image and signal processing, *Bio. Med. Eng. Inf.,* 2020, 1004–1009. https://doi.org/10.1109/cisp-bmei51763.2020.9263602
6. Y. Cheng, D. M. Titterington, Neural networks: A review from a statistical perspective, *Stat. Sci.*, **9** (1994), 2–45. http://www.jstor.org/stable/2246275.

7. S. H. Kim, I. Han, Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index, *Expert Syst. Appl.*, **19** (2000), 125–132. https://doi.org/10.1016/S0957-4174(00)00027-0
8. S. Lahmiri, *Modeling Stock Market Industrial Sectors as Dynamic Systems and Forecasting*, In: Encyclopedia of Information Science and Technology, Third Edition, IGI Global, 2015, 3818–3830. https://doi.org/10.4018/978-1-4666-5888-2.ch376
9. Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, *P. IEEE*, **86** (1998), 2278–2324. https://doi.org/10.1109/5.726791
10. P. Y. Simard, D. Steinkraus, J. C. Platt, Best practices for convolutional neural networks applied to visual document analysis, *In: Icdarm* **3** (2003), No. 2003. https://doi.org/10.1109/icdar.2003.1227801
11. X. Glorot, A. Bordes, Y. Bengio, *Deep sparse rectifier neural networks, In: Proceedings of the fourteenth international conference on artificial intelligence and statistics*, JMLR Workshop and Conference Proceedings, 2011, 315–323.
12. M. Daoust, J.Bégin, C. Gagné, *Data augmentation using conditional generative adversarial networks for the detection of cyberbullying, In: Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining,* IEEE, 2016, 615–618. https://doi.org/10.1109/asonam.2016.7752342
13. A. Wong, C. Leung, *A review on data augmentation techniques for deep learning, In: Proceedings of the 31st International Conference on Neural Information Processing Systems,* 2018, 2234–2244.
14. H. Yao, S. Zhao, Z. Gao, Z. Xue, B. Song, F. Li, et al., Data-driven analysis on the subbase strain prediction: A deep data augmentation-based study, *Transp. Geotech.*, **40** (2023), 100957. https://doi.org/10.1016/j.trgeo.2023.100957
15. J. Yoon, J. Jordon, M. van der Schaar, TimeGAN: Preprocessing raw data for time series generation with generative adversarial networks, *Proceedings of the 36th International Conference on Machine Learning*, **48** (2019), 7272–7281.
16. A. Rasheed, O. San, T. Kvamsdal, Digital twin: Values, challenges and enablers from a modeling perspective, *Ieee Access*, **8** (2020), 21980–22012. https://doi.org/10.1109/access.2020.2970143
17. J. Jeon, J. Kim, H. Song, S. Cho, N. Park, GT-GAN: General purpose time series synthesis with generative adversarial networks, *Adv. Neural Inf. Process. Syst.*, **35** (2022), 36999–37010.
18. P., Chlap, H. Min, N. Vandenberg, J. Dowling, L. Holloway, A. Haworth, A review of medical image data augmentation techniques for deep learning applications, *J. Med. Imag. Radiat. On.*, **65** (2021), 545–563. https://doi.org/10.1111/1754-9485.13261
19. H. Naveed, *Survey: image mixing and deleting for data augmentation*, arXiv preprint arXiv: 2106.07085, 2021. https://doi.org/10.48550/arXiv.2106.07085
20. S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, et al., A survey of data augmentation approaches for nlp, arXiv preprint arXiv:2105.03075, 2021. https://doi.org/10.18653/v1/2021.findings-acl.84
21. Q. Wen, L. Sun, F. Yang, X. Song, J. Gao, X. Wang, et al., *Time series data augmentation for deep learning: A survey*. arXiv preprint arXiv:2002.12478, 2020. https://doi.org/10.48550/arXiv.2002.12478
22. G. García-Molina, E. Gómez-Sánchez, A. García-Sánchez, Data Augmentation by Imputation Techniques in Time Series: Application to the Spanish Electricity Market, *Processes*, **7** (2021), 958. https://doi.org/10.3390/pr7120958.

23. A. Le Guennec, S. Malinowski, R. Tavenard, *Data augmentation for time series classification using convolutional neural networks*, In: ECML/PKDD workshop on advanced analytics and learning on temporal data, 2016. https://doi.org/10.1007/978-3-030-91445-5
24. D. Salinas, S. Mehrotra, S. Mohan, *DeepAR: Probabilistic forecasting with autoregressive recurrent networks*, arXiv preprint arXiv:1704.04110, 2020. https://doi.org/10.1016/j.ijforecast.2019.07.001
25. M. S. Gashler, S. C. Ashmore, Training deep fourier neural networks to fit time-series data. In: Intelligent Computing in Bioinformatics: 10th International Conference, ICIC 2014, Taiyuan, China, August 3–6, 2014, *Proceedings* **10** (2014), 48–55. Springer International Publishing. https://doi.org/10.1007/978-3-319-09330-7_7
26. H. Kim, J. Kim, S. Oh, *Time series prediction with Monte Carlo tree search and online learning*. In 2017 IEEE International Conference on Big Data (Big Data), 2017, 3495–3500. https://doi.org/10.1109/bigdata47090.2019.9006276
27. Gao, C., Zhang, N., Li, Y., Bian, F., & Wan, H.. Self-attention-based time-variant neural networks for multi-step time series forecasting. *Neural Computing and Applications*, **34(11)** (2022), 8737-8754. https://doi.org/10.1007/s00521-021-06871-1
28. Li, Z., Ma, C., Shi, X., Zhang, D., Li, W., & Wu, L. Tsa-gan: A robust generative adversarial networks for time series augmentation. In 2021 International Joint Conference on Neural Networks (IJCNN) 2021, 1–8. https://doi.org/10.1109/IJCNN52387.2021.9534001
29. X., Tan, X. Sun, W. Chen, B. Du, J. Ye, L. Sun, Investigation on the data augmentation using machine learning algorithms in structural health monitoring information, *Struct. Health Monit.*, **20** (2021), 2054–2068. https://doi.org/10.1177/1475921721996238
30. D. P. Kingma, M. Welling, *Auto-encoding variational bayes*, arXiv preprint arXiv:1312.6114. 2013. https://doi.org/10.48550/arXiv.1312.6114
31. L. Deng, Deep learning: from speech recognition to language and multimodal processing, APSIPA *Trans. Signal*, **5** (2016). https://doi.org/10.1017/ATSIP.2015.22
32. I. Goodfellow, Y. Bengio, A. Courville, *Deep learning*, MIT press, 2016.
33. P. Isola, J. Y. Zhu, T. Zhou, A. A. Efros, *Image-to-image translation with conditional adversarial networks*, In Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, 1125–1134. https://doi.org/10.1109/cvpr.2017.632
34. J. Cheng, Y. Yang, X. Tang, N. Xiong, Y. Zhang, F. Lei, Generative Adversarial Networks: A Literature Review, *KSII T. Internet Inf.*, **14** (2020), 4625–4647. https://doi.org/10.3837/tiis.2020.12.001
35. S. Hochreiter, J. Schmidhuber, Long short-term memory, *Neural Comput.*, **9** (1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
36. A. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, et al., *A generative model for raw audio*. arXiv preprint arXiv:1609.03499, 2016. https://doi.org/10.48550/arXiv.1609.03499
37. L. Dinh, J. Sohl-Dickstein, S. Bengio, *Density estimation using Real NVP*. arXiv preprint arXiv:1605.08803, 2014. https://doi.org/10.48550/arXiv.1605.08803
38. S. Suradhaniwar, S. Kar, S. S. Durbha, A. Jagarlapudi, Time series forecasting of univariate agrometeorological data: a comparative performance evaluation via one-step and multi-step ahead forecasting strategies, *Sensors*, **21** (2021), 2430. https://doi.org/10.3390/s21072430
39. J. H. Poynting, A comparison of the fluctuations in the price of wheat and in cotton and silk imports into Great Britain, *J. Roy. Stat. Soc.*, **47** (1884), 34–74. https://doi.org/10.2307/2979211

40. R. H. Hooker, Correlation of the marriage-rate with trade, *J. Roy. Stat. Soc.*, **64** (1901), 485–492.
41. J. Spencer, On the graduation of the rates of sickness and mortality presented by the experience of the Manchester Unity of Oddfellows during the period 1893–97, *J. Institute Actuaries*, **38** (1904), 334–343. https://doi.org/10.1017/s0020268100008076
42. R. G. Brown, *Smoothing, forecasting and prediction of discrete time series*, Prentice-Hall, 1963.
43. D. E. Rumelhart, G. E. Hinton, R. J. Williams, Learning representations by back-propagating errors, *Nature*, **323** (1986), 533–536. https://doi.org/10.1038/323533a0
44. G. Zhang, B. E. Patuwo, M. Y. Hu, Forecasting with artificial neural networks: The state of the art, *Int. J. Forecasting*, **14** (1998), 35–62. https://doi.org/10.1016/S0169-2070(97)00044-7
45. K. Hornik, M. Stinchcombe, H. White, Multilayer feedforward networks are universal approximators, *Neural Network.*, **2** (1989), 359–366. https://doi.org/10.1016/0893-6080(89)90020-8
46. P. J. Werbos, Generalization of backpropagation with application to a recurrent gas market model, *Neural Network.*, **1** (1988), 339–356. https://doi.org/10.1016/0893-6080(88)90007-x
47. T. Hill, M. O'Connor, W. Remus, Neural network models for time series forecasts, *Manag. Sci.*, **42** (1996), 1082–1092. https://doi.org/10.1287/mnsc.42.7.1082
48. S. Siami-Namini, A. S. Namin, *Forecasting economics and financial time series: ARIMA vs. LSTM*. arXiv preprint arXiv:1803.06386, 2018. https://doi.org/10.1109/icmla.2018.00227
49. V. Ravi, D. Pradeepkumar, K. Deb, Financial time series prediction using hybrids of chaos theory, multi-layer perceptron and multi-objective evolutionary algorithms, *Swarm Evol. Comput.*, **36** (2017), 136–149. https://doi.org/10.1016/j.swevo.2017.05.003
50. A. Zameer, A. Khan, S. G. Javed, Machine learning based short term wind power prediction using a hybrid learning model, *Comput, Electr, Eng,*, **45** (2015), 122–133. https://doi.org/10.1016/j.compeleceng.2014.07.009
51. M. Jiang, L. Jia, Z. Chen, W. Chen, The two-stage machine learning ensemble models for stock price prediction by combining mode decomposition, extreme learning machine and improved harmony search algorithm, *Ann. Oper. Res.,* 309 (2022), 533–585.
52. P. Du, J. Wang, W. Yang, T. Niu, A novel hybrid model for short-term wind power forecasting, *Appl. Soft Comput.*, **80** (2019), 93–106. https://doi.org/10.1016/j.asoc.2019.03.035
53. A. Lazcano, P. J. Herrera, M. A. Monge, Combined model based on recurrent neural networks and graph convolutional networks for financial time series forecasting, *Mathematics*, **11** (2023), 224. https://doi.org/10.3390/math11010224
54. S. X. Lv, L. Wang, Multivariate wind speed forecasting based on multi-objective feature selection approach and hybrid deep learning model, *Energy*, **263** (2023), 126100. https://doi.org/10.1016/j.energy.2022.126100
55. F. Wang, Z. Zhang, C. Liu, Y. Yu, S. Pang, N. Duić, et al., Generative adversarial networks and convolutional neural networks based weather classification model for day ahead short-term photovoltaic power forecasting, *Energ. Convers. Manage.*, **181** (2019), 443–462. https://doi.org/10.1016/j.enconman.2018.11.074
56. K. M. Rashid, J. Louis, Times-series data augmentation and deep learning for construction equipment activity recognition, *Adv. Eng. Inform.*, **42** (2019), 100944. https://doi.org/10.1016/j.aei.2019.100944
57. Y. Luo, X. Cai, Y. Zhang, J. Xu, Multivariate time series imputation with generative adversarial networks, *Adv. Neural Inform. Proces. Syst.*, **31** (2018).

58. Q. Wen, L. Sun, F. Yang, X. Song, J. Gao, X. Wang, et al., Time series data augmentation for deep learning: A survey, arXiv preprint arXiv:2002.12478, 2020. https://doi.org/10.48550/arXiv.2002.12478
59. C. Shorten, T. M. Khoshgoftaar, A survey on image data augmentation for deep learning, *J. Big Data*, **6** (2019), 1–48. https://doi.org/10.1186/s40537-019-0197-0
60. P. H. Hsieh, P. H. Liao, A comparative study of stock price forecasting models, *J. Appl. Math.*, **111** (2019). https://doi.org/10.1155/2019/8681410
61. T. C.Tung, S. H.Yen, T. Y. Huang, C. P.Chen, Short-Term electric load forecasting using stacked extreme learning machine with clustering technique, *Energies*, **13** (2020), 3977. https://doi.org/10.3390/en13153977
62. K. Bandara, H. Hewamalage, Y. H. Liu, Y. Kang, C. Bergmeir, Improving the accuracy of global forecasting models using time series data augmentation, *Pattern Recogn.*, **120** (2021), 108148. https://doi.org/10.1016/j.patcog.2021.108148
63. M. F. Asem, M. M. Abogameel, N. Almujally, A. H. Alkashan, *Comparative study of interpolation methods for time series data augmentation*, Proceedings of the 2021 International Conference on High Performance Computing & Simulation, 2021, 110–115.

AIMS Press