Complexity

Complexity in Forecasting and Predictive Models

Lead Guest Editor: Jose L. Salmeron
Guest Editors: Marisol B. Correia and Pedro Palos

Complexity in Forecasting and Predictive Models

Complexity in Forecasting and Predictive Models

Lead Guest Editor: Jose L. Salmeron
Guest Editors: Marisol B. Correia and Pedro Palos

Editorial Board

Contents

Hindawi
Complexity
Volume 2019, Article ID 8160659, 3 pages
https://doi.org/10.1155/2019/8160659

Editorial

Complexity in Forecasting and Predictive Models

Jose L. Salmeron,[1] **Marisol B. Correia**,[2] **and Pedro R. Palos-Sanchez**[3]

[1] *Universidad Pablo de Olavide de Sevilla, Spain*
[2] *ESGHT, University of Algarve & CiTUR - Centre for Tourism Research, Development and Innovation & CEG-IST, Instituto Superior Técnico, Universidade de Lisboa, Portugal*
[3] *International University of La Rioja, Spain*

Correspondence should be addressed to Jose L. Salmeron; salmeron@acm.org

Received 20 May 2019; Accepted 21 May 2019; Published 10 June 2019

1. Introduction

The challenge of this special issue has been to know the state of the problem related to forecasting modeling and the creation of a model to forecast the future behavior that supports decision making by supporting real-world applications.

This issue has been highlighted by the quality of its research work on the critical importance of advanced analytical methods, such as neural networks, soft computing, evolutionary algorithms, chaotic models, cellular automata, agent-based models, and finite mixture minimum squares (FIMIX-PLS)

Mainly, all the papers are focused on triggering a substantive discussion on how the model predictions can face the challenges around the complexity field that lie ahead. These works help to better understand the new trends in computing and statistical techniques that allow us to make better forecasts. Complexity plays a prominent role in these trends, given the increasing variety and changing data flows, forcing academics to adopt innovative and hybrid methods.

The papers address the recent advances in methodological issues and practices related to the complexity of forecasting and models. All of them are increasingly focused on heterogeneous sources.

Among these techniques, FIMIX-PLS has been applied to study unobserved heterogeneity [1]. This technique was extended by Sarstedt et al. [2]. FIMIX-PLS is the first latent class approach for PLS-SEM [3] and is an exploration tool that allows obtaining the appropriate number of segments in which the sample will be divided. In this sense, the FIMIX-PLS technique allowed making decisions regarding the number of segments, based on pragmatic reasons and considering practical issues related to current research [4].

FIMIX-PLS calculates the probabilities of belonging to a given segment in which each observation is adjusted to the predetermined numbers of segments by means of the estimation of separate linear regression functions and the belonging of objects corresponding to several segments. Different cases are assigned to the segment with greater probability.

2. Static Control Model

A static control model based on Wavelet Transform Weighted Twin Support Vector Regression (WTWTSVR) is proposed in the paper "End-Point Static Control of Basic Oxygen Furnace (BOF) Steelmaking Based on Wavelet Transform Weighted Twin Support Vector Regression" by C. Gao et al. The first approach of the authors was to add a new weighted matrix and coefficient vector into the objective functions of Twin Support Vector Regression (TSVR) to improve the performance of the algorithm, and then a static control model was established based on WTWTSVR and 220 samples in real plant, which consists of prediction models, control models, regulating units, controller, and Basic Oxygen Furnace (BOF).

The results indicated that the control error bound with 800 Nm^3 in the oxygen blowing volume and 5.5 tons in the weight of auxiliary materials can achieve a hit rate of 90% and 88%, respectively. In conclusion, the proposed model can provide a significant reference for real BOF applications

and can be extended to the prediction and control of other industry applications.

3. Hybrid Models

In the paper "Development of Multidecomposition Hybrid Model for Hydrological Time Series Analysis" by H. M. Nazir et al., two hybrid models were developed to improve the prediction precision of hydrological time series data based on the principal of three stages as denoising, decomposition, and decomposed component prediction and summation. The performance of the proposed models was compared with the traditional single-stage model (without denoised and decomposed), with the hybrid two-stage model (with denoised), and with the existing three-stage hybrid model (with denoised and decomposition). Three evaluation measures were used to assess the prediction accuracy of all models (Mean Relative Error (MRE), Mean Absolute Error (MAE), and Mean Square Error (MSE)). The proposed architecture was applied on daily rivers inflow time series data of Indus Basin System and the results showed that the three-stage hybrid models had shown improvement in prediction accuracy with minimum MRE, MAE, and MSE for all cases and that the accuracy of prediction was improved by reducing the complexity of hydrological time series data by incorporating the denoising and decomposition.

4. Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN)

Y. Zhou et al. proposed an approach that integrates complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) and extreme gradient boosting (XGBOOST), the so-called CEEMDAN-XGBOOST, for forecasting crude oil prices in the paper "A CEEMDAN and XGBOOST-Based Approach to Forecast Crude Oil Prices". To demonstrate the performance of the proposed approach, the authors conducted extensive experiments on the West Texas Intermediate (WTI) crude oil prices. Finally, the experimental results showed that the proposed CEEMDAN-XGBOOST outperforms some state-of-the-art models in terms of several evaluation metrics.

5. Machine Learning (ML) Algorithms

The ability to handle large amounts of data incrementally and efficiently is indispensable for Machine Learning (ML) algorithms. The paper "An Incremental Learning Ensemble Strategy for Industrial Process Soft Sensors" by H. Tian et al. addressed an Incremental Learning Ensemble Strategy (ILES) that incorporates incremental learning to extract information efficiently from constantly incoming data. The ILES aggregates multiple sublearning machines by different weights for better accuracy.

The weight updating rules were designed by considering the prediction accuracy of submachines with new arrived data, so that the update can fit the data change and obtain new information efficiently. The sizing percentage soft sensor model was established to learn the information from the production data in the sizing of industrial processes and to test the performance of ILES, where the Extreme Learning Machine (ELM) was selected as the sublearning machine. The results of the experiments demonstrated that the soft sensor model based on the ILES has the best accuracy and ability of online updating.

6. Bass Innovation Diffusion Model

M. L. Bertotti and G. Modanese in "The Bass Diffusion Model on Finite Barabasi-Albert Networks" used a heterogeneous mean-field network formulation of the Bass innovation diffusion model and exact results by Fotouhi and Rabbat [5] on the degree correlations of Barabasi-Albert (BA) networks to compute the times of the diffusion peak and to compare them with those on scale-free networks, which have the same scale-free exponent but different assortativity properties.

The authors compared their results with those obtained by D'Agostino *et al.* [6] for the SIS epidemic model with the spectral method applied to adjacency matrices and turned out that diffusion times on finite BA networks were at a minimum. They believe this may be due to a specific property of these networks, e.g. whereas the value of the assortativity coefficient is close to zero, the networks look disassortative if a bounded range of degrees is solely considered, including the smallest ones, and slightly assortative on the range of the higher degrees.

Finally, the authors found that if the trickle-down character of the diffusion process is enhanced by a larger initial stimulus on the hubs (via a inhomogeneous linear term in the Bass model), the relative difference between the diffusion times for BA networks and uncorrelated networks is even larger.

7. Forecast Copper Treatment Charges (TC)/Refining Charges (RC)

In order to overcome the lack of research about the price at which mines sell copper concentrate to smelters, the paper "Looking for Accurate Forecasting of Copper TC/RC Benchmark Levels" by F. J. Díaz-Borrego et al. provides a tool to forecast copper Treatment Charges (TC)/Refining Charges (RC) annual benchmark levels, in which a three-model comparison was made by contrasting different measures of error.

The results indicated that the Linear Exponential Smoothing (LES) model is the one that has the best predictive capacity to explain the evolution of TC/RC in both the long and the short term. Finally, the authors believe this suggests a certain dependency on the previous levels of TC/RC, as well as the potential existence of cyclical patterns in them and that this model enables a more precise estimation of copper TC/RC levels, which is useful for smelters and mining companies.

8. Ant Colony Optimization Algorithm

F. Liu et al. in "Prediction of Ammunition Storage Reliability Based on Improved Ant Colony Algorithm and BP Neural Network" proposed an Improved Ant Colony Optimization (IACO) algorithm based on a three-stage and a Back Propagation (BP) Neural Network algorithm to predict ammunition failure numbers, where the main affecting factors of ammunition include temperature, humidity, and storage period and where the reliability of ammunition storage is obtained indirectly by failure numbers. Experimental results show that IACO-BP algorithm has better accuracy and stability in ammunition storage reliability prediction than BP network, Particle Swarm Optimization-Back Propagation (PSO-BP), and ACO-BP algorithms.

9. Finite Mixture Partial Least Squares (FIMIX-PLS)

Finally, the paper "Green Start-Ups' Attitudes towards Nature When Complying with the Corporate Law" by R. Robina-Ramírez et al. examined how Spanish green start-ups develop improved attitudes towards nature having in mind the framework of the new Spanish Criminal Code in relation to corporate compliance. Smart Partial Least Squares (PLS) Path Modelling was used to build an interaction model among variables and unobserved heterogeneity was analysed using Finite Mixture Partial Least Squares (FIMIX-PLS).

The model reveals a strong predictive power (R^2 =77.9%) with a sampling of one hundred and fifty-two Spanish start-ups. This test is performed in a set of four stages. In the first one, it executes FIMIX, and in a second it calculates the number of optimal segments. In the third one, we discussed the latent variables that justify these segments, to finally estimate the model and segments [7].

The results allow concluding that Spanish start-ups should implement effective monitoring systems and new organisational standards, in order to develop surveillance measures to avoid unexpected sanctions.

10. Conclusion

An overview of the 8 selected papers for this special issue has been presented, reflecting the most recent progress in the research field. We hope this special issue can further stimulate interest in the critical importance of advanced analytical methods, such as neural networks, soft computing, evolutionary algorithms, chaotic models, cellular automata, agent-based models, and finite mixture minimum squares (FIMIX-PLS). More research results and practical applications on advanced analytical methods are expected in the future.

Conflicts of Interest

The guest editor team does not have any conflicts of interest or private agreements with companies.

Jose L. Salmeron
Marisol B. Correia
Pedro R. Palos-Sanchez

References

[1] C. Hahn, M. D. Johnson, A. Herrmann, and F. Huber, "Capturing customer heterogeneity using a finite mixture PLS approach," *Schmalenbach Business Review*, vol. 54, no. 3, pp. 243–269, 2002 (Russian).

[2] M. Sarstedt, J.-M. Becker, C. M. Ringle, and M. Schwaiger, "Uncovering and treating unobserved heterogeneity with FIMIX-PLS: which model selection criterion provides an appropriate number of segments?" *Schmalenbach Business Review*, vol. 63, no. 1, pp. 34–62, 2011.

[3] M. Sarstedt, "A review of recent approaches for capturing heterogeneity in partial least squares path modelling," *Journal of Modelling in Management*, vol. 3, no. 2, pp. 140–161, 2008.

[4] M. Sarstedt, M. Schwaiger, and C. M. Ringle, "Do we fully understand the critical success factors of customer satisfaction with industrial goods? - extending festge and schwaigers model to account for unobserved heterogeneity," *Journal of Business Market Management*, vol. 3, no. 3, pp. 185–206, 2009.

[5] B. Fotouhi and M. G. Rabbat, "Degree correlation in scale-free graphs," *The European Physical Journal B*, vol. 86, no. 12, article no 510, 2013.

[6] G. D'Agostino, A. Scala, V. Zlatić, and G. Caldarelli, "Robustness and assortativity for di?usion-like processes in scale-free networks," *EPL (Europhysics Letters)*, vol. 97, no. 6, article no 68006, 2012.

[7] P. Palos-Sanchez, F. Martin-Velicia, and J. R. Saura, "Complexity in the acceptance of sustainable search engines on the internet: an analysis of unobserved heterogeneity with FIMIX-PLS," *Complexity*, Article ID 6561417, 19 pages, 2018.

Hindawi
Complexity
Volume 2019, Article ID 5353296, 12 pages
https://doi.org/10.1155/2019/5353296

Research Article

An Incremental Learning Ensemble Strategy for Industrial Process Soft Sensors

Huixin Tian,[1,2,3] **Minwei Shuai**,[1] **Kun Li**,[4] **and Xiao Peng**[1]

[1] *School of Electrical Engineering & Automation and Key Laboratory of Advanced Electrical Engineering and Energy Technology, Tianjin Polytechnic University, Tianjin, China*
[2] *State Key Laboratory of Process Automation in Mining & Metallurgy, Beijing 100160, China*
[3] *Beijing Key Laboratory of Process Automation in Mining & Metallurgy Research Fund Project, Beijing 100160, China*
[4] *School of Economics and Management, Tianjin Polytechnic University, Tianjin, China*

Correspondence should be addressed to Kun Li; lk_neu@163.com

Received 13 November 2018; Accepted 25 March 2019; Published 2 May 2019

Guest Editor: Marisol B. Correia

With the continuous improvement of automation in industrial production, industrial process data tends to arrive continuously in many cases. The ability to handle large amounts of data incrementally and efficiently is indispensable for modern machine learning (ML) algorithms. According to the characteristics of industrial production process, we address an ILES (incremental learning ensemble strategy) that incorporates incremental learning to extract information efficiently from constantly incoming data. The ILES aggregates multiple sublearning machines by different weights for better accuracy. When new data set arrives, a new submachine will be trained and aggregated into ensemble soft sensor model according to its weight. The other submachines' weights will be updated at the same time. Then a new updated soft sensor ensemble model can be obtained. The weight updating rules are designed by considering the prediction accuracy of submachines with new arrived data. So the update can fit the data change and obtain new information efficiently. The sizing percentage soft sensor model is established to learn the information from the production data in the sizing of industrial processes and to test the performance of ILES, where the ELM (Extreme Learning Machine) is selected as the sublearning machine. The comparison is done among new method, single ELM, AdaBoost.R ELM, and OS-ELM, and the test of the extensions is done with three test functions. The results of the experiments demonstrate that the soft sensor model based on the ILES has the best accuracy and ability of online updating.

1. Introduction

During industrial processes, plants are usually heavily instrumented with a large number of sensors for process monitoring and control. However, there are still many process parameters that cannot be measured accurately because of high temperature, high pressure, complex physical and chemical reactions and large delays, etc. Soft sensor technology provides an effective way to solve these problems. The original and still the most dominant application area of soft sensors is the prediction of process variables, which can be determined either at low sampling rates or through off-line analysis only. Because these variables are often related to the process product quality, they are very important for process control and quality management. Additionally, the soft sensor application field usually refers to online prediction during the process of production.

Currently, with the continuous improvement of automation in industrial production, large amounts of industrial process data can be measured, collected, and stored automatically. It can provide strong support for the establishment of data-driven soft sensor models. Meanwhile, with the rapid development and wide application of big data technology, soft sensor technology has already been used widely and plays an essential role in the development of industrial process detection and control systems in industrial production. Artificial intelligence and machine learning, as the important core technologies, are getting increasingly more attention. The traditional machine learning algorithm generally refers to the single learning machine model that

is trained by training sets, and then the unknown samples will be predicted based on this model. However, the single learning machine models have to face defects that cannot be overcome by themselves, such as unsatisfactory accuracy and generalization performance, especially for complex industrial processes. Specifically, in the supervised machine learning approach, the model's hypothesis is produced to predict the new incoming instances by using predefined label instances. When the multiple hypotheses that support the final decision are aggregated together, it is called ensemble learning. Compared with the single learning machine model, the ensemble learning technique is beneficial for improving quality and accuracy. Therefore, increasingly more researchers are studying how to improve the speed, accuracy and generalization performance of integrated algorithms instead of developing strong learning machines.

Ensemble algorithms were originally developed for solving binary classification problems [1], and then AdaBoost.M1 and AdaBoost.M2 were proposed by Freund and Schapire [2] for solving multiclassification problems. Thus far, there are many different versions of boosting algorithms for solving classification problems [3–7], such as boosting by filtering and boosting by subsampling. However, for regression problems, it is not possible to predict the output exactly as that in classification. To solve regression problems using ensemble techniques, Freund and Schapire [2] extended AdaBoost.M2 to AdaBoost.R, which projects the regression sample into a classification data set. Drucker [8] proposed the AdaBoost.R2 algorithm, which is an ad hoc modification of AdaBoost.R. Avnimelech and Intrator [9] extended the boosting algorithm for regression problems by introducing the notion of weak and strong learning as well as an appropriate equivalence theorem between the two. Feely [10] proposed BEM (big error margin) boosting method, which is quite similar to the AdaBoost.R2. In BEM, the prediction error is compared with the preset threshold value, BEM, and the corresponding example is classified as either well or poorly predicted. Shrestha and Solomatine [11] proposed an AdaBoost.RT, with the idea of filtering out the examples with relative estimation errors that are higher than the preset threshold value. However, the value of the threshold is difficult to set without experience. For solving this problem, Tian and Mao [12] present a modified AdaBoost.RT algorithm that can adaptively modify the threshold value according to the change in RMSE. Although ensemble algorithms have the ability to enhance the accuracy of soft sensors, they are still at a loss for the further information in the new coming data.

In recent years, with the rapid growth of data size, a fresh research perspective has arisen to face the large amount of unknown important information contained in incoming data streams in various fields. How can we obtain methods that can quickly and efficiently extract information from constantly incoming data? The batch learning is meaningless, and the algorithm needs to have the capability of real-time processing because of the demand of real-time updated data in industrial processes. Incremental learning idea is helpful to solve the above problem. If a learning machine has this kind of idea, it can learn new knowledge from new data sets and can retain old knowledge without accessing the old data set. Thus, the incremental learning strategy can profoundly increase the processing speed of new data while also saving the computer's storage space. The ensemble learning methods can be improved by combining the characteristics of the ensemble strategy and incremental learning. It is an effective and suitable way to solve the problem of stream data mining [13–15]. Learn++ is a representative ensemble algorithm with the ability of incremental learning. This algorithm is designed by Polikar et al. based on AdaBoost and supervised learning [16]. In Learn ++, the new data is also assigned sample weights, which update according to the results of classification at each iteration. Then, a newly trained weak classifier is added to the ensemble classifier. Based on the Learn ++ CDS algorithm, G Ditzler and Polikar proposed Learn ++ NIE [17] to improve the effect of classification on a few categories. Most of research of incremental learning and ensemble algorithm focus on the classification field, while the research in the field of regression is still less. Meanwhile the limitation of ensemble approaches is that they cannot address the essential problem of incremental learning well, what the essential problem is accumulating experience over time then adapting and using it to facilitate future learning process [18–20].

For the process of industrial production, increasingly more intelligent methods are used in soft sensors with the fast development of artificial intelligence. However, the practical applications of soft sensors in industrial production are not good. The common shortages of soft sensors are unsatisfactory, unstable prediction accuracy, and poor online updating abilities. It is difficult to meet a variety of changes in the process of industrial production. Therefore, in this paper, we mainly focus on how to add the incremental learning capability to the ensemble soft sensor modeling method and hopefully provide useful suggestions to enhance both the generation and online application abilities of soft sensors for industrial process. Aiming at the demands of soft sensors for industrial applications, a new detection strategy is proposed with multiple learning machines ensembles to improve the accuracy of the soft sensors based on intelligent algorithms. Additionally, in practical production applications, acquisition of information in new production data is expensive and time consuming. Consequently, it is necessary to update the soft sensor in an incremental fashion to accommodate new data without compromising the performance on old data. Practically, in most traditional intelligent prediction models for industrial process, the updates are often neglected. Some models use traditional updating methods that retrain the models by using all production data or using the updated data and forgo the old data. This kind of methods is not good enough because some good performances have to be lost to learn new information [21]. Against this background, we present a new incremental learning ensemble strategy with a better incremental learning ability to establish the soft sensor model, which can learn additional information from new data and preserve previously acquired knowledge. The update does not require the original data that was used to train the existing old model.

In the rest of this paper, we first describe the details of the incremental learning ensemble strategy (ILES), which

involves the strategy of updating the weights, the ensemble strategy, and the strategy of incremental learning for real-time updating. Then, we design experiments to test the performance of the ILES for industrial process soft sensors. The sizing percentage of the soft sensor model is built by the ILES in the sizing production process. The parameters of the ILES are discussed. We also compare the performance of the ILES to those of other methods. To verify the universal use of the new algorithm, three test functions are used to test the improvement on the predictive performance of the ILES. Finally, we summarize our conclusions and highlight future research directions.

2. The Incremental Learning Ensemble Strategy

The industrial process needs soft sensors with good accuracy and online updating performance. Here, we focus on incorporating the incremental learning idea into the ensemble regression strategy to achieve better performance of soft sensors. A new ensemble strategy called ILES for industrial process soft sensors that combines the ensemble strategy with the incremental learning idea is proposed. The ILES has the ability to enhance soft sensors' accuracy by aggregating the multiple sublearning machines according to their training errors and prediction errors. Additionally, during the iteration process, incremental learning is added to obtain the information from new data by updating the weights. It is beneficial to enhance the real-time updating ability of industrial process online soft sensors. The details of the ILES are described as shown in Algorithm 1.

2.1. Strategy of Updating the Weight. In each iteration of k ($k = 1, 2, \ldots, K$), the initial $D_1(i) = 1/m(k)$ is distributed to each sample (x_i, y_i) with the same values. This means that the samples have the same chance to be included in training dataset or tested dataset at the beginning. In the subsequent iterations, the weight will be calculated as $D_t(i) = w_t / \sum_{i=1}^{m} w_t(i)$ for every sublearning machine (in each iteration of t). In contrast to the traditional AdaBoost.R, here, the testing subdataset is added to test the learning performance in each iteration. It is useful to ensure the generalization performance of the ensemble soft sensors. Then, the distribution will be changed according to the training and testing errors at the end of each iteration. Here, the training subdataset TR_t and the testing subdataset TE_t will be randomly chosen according to D_t (for example, by using the roulette method). The sublearning machine is trained by TR_t, and a hypothesized soft sensor $f_t : x \longrightarrow y$ will be obtained. Then, the training error and testing error of f_t can be calculated as follows:

$$ARE_t(i) = \left| \frac{f(x_i) - y_i}{y_i} \right| \tag{1}$$

The error rate of f_t on $S_k = TR_t + TE_t$ is defined as follows:

$$\varepsilon_t = \sum_{ARE_t(i) > \delta} D_t(i) \tag{2}$$

If $\varepsilon_t > \varphi$, the submodel is regarded as an unqualified and suspicious model. The hypothesis f_t is given up. Otherwise, the power coefficient is calculated as $\beta_t = \varepsilon_t^n$ (e.g., linear, square, or cubic). Here, φ ($0 < \varphi < 1$) is the coefficient of determination. After t ($t = 1, 2, \ldots, T_k$) iterations, the composite hypothesis F_k can be obtained according to the t hypothesized soft sensors (sublearning machines) $f_1, f_2, \ldots, f_t$. The training subdataset error, the testing subdataset error, and the error rate of F_k are calculated similarly to those of f_t. In the same way, if $E_k > \varphi$, the hypothesis F_k is given up. At the end of the iterations, according to the error rate E_t, the weight is updated as follows:

$$w_{t+1} = w_t \times \begin{cases} B_k & ARE_k(i) < \delta \\ 1 & \text{otherwise} \end{cases} \tag{3}$$

where $B_k = E_k/(1 - E_k)$. In the next iteration, the TR_t and TE_t will be chosen again according to the new distribution, which is calculated by the new weight w_{t+1}. During the above process of iterations, the updating of the weights depends on the training and testing performance of the sublearning machines with different data. Therefore, the data with large errors will have a larger distribution for the difficult learning. It means that the "difficult" data will have more chances to be trained until the information in the data is obtained. Conversely, the sublearning machines or hypothesized soft sensors are reserved selectively based on their performance. Therefore, the final hypothesized soft sensors are well qualified to aggregate the composite hypothesis. This strategy is very effective for improving the accuracy of ensemble soft sensors.

2.2. Strategy of Ensemble with Incremental Learning. Aiming at the needs of real-time updates, the incremental learning strategy is integrated into the ensemble process. First, the subdatasets $S_k = [(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)]$, $k = 1, 2, \ldots, K$ are selected randomly from the dataset. In each iteration $k = 1, 2, \ldots, K$, the sublearning machines are trained and tested. Therefore, for each subdataset, when the inner loop ($t = 1, 2, \ldots, T_k$) is finished, the T_k hypothesized soft sensors $f_1, f_2, \ldots, f_{T_k}$ are generated. An ensemble soft sensor is obtained based on the combined outputs of the individual hypotheses, which constitute the composite hypothesis F_k.

$$F_k(x) = \frac{\sum_t \left(\lg\left(1/\beta_t\right)\right) f_t(x)}{\sum_t \left(\lg\left(1/\beta_t\right)\right)} \tag{4}$$

Here, the better hypotheses will be aggregated with larger chances. Therefore, the best performance of ensemble soft sensors is ensured based on these sublearning machines. Then, the training subdataset error and testing subdataset error of F_k can be calculated similarly to the error of f_t.

$$ARE_t(i) = \left| \frac{F_t(x_i) - y_i}{y_i} \right| \tag{5}$$

The error rate of F_k on $S_k = TR_t + TE_t$ is defined as follows:

$$E_k = \sum_{ARE_k(i) > \delta} D_t(i) \tag{6}$$

Input

(i) S_k sub datasets are drawn from original data set. Here, $k = 1, 2, \ldots, K$, $S_k = [(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)]$.
(ii) The number of sub learning machines is T_k.
(iii) The coefficients of determination are δ and φ.

For $k = 1, 2, \ldots, K$

Initialize $D_1(i) = 1/m(k)$, $\forall i = 1, 2, \ldots, K$. Here, $m(k)$, is the amount of data.

For $t = 1, 2, \ldots, T_k$

(1) Calculate $D_t(i) = w_t / \sum_{i=1}^{m} w_t(i)$. D_t is a distribution, and $w_t(i)$ is the weight of (x_i, y_i).
(2) Randomly choose the training sub dataset TR_t and the testing sub dataset TE_t according to D_t.
(3) The sub learning machine is trained by TR_t to obtain a soft sensor model $f_t : x \longrightarrow y$.
(4) Calculate the error of f_t using TR_t and TE_t:

$$ARE_t(i) = \left| \frac{f(x_i) - y_i}{y_i} \right|.$$

(5) Calculate the error rate $\varepsilon_t = \sum_{ARE_t(i)>\delta} D_t(i)$. If $\varepsilon_t > \varphi$, give up f_t, and return to step (2).
(6) Calculate $\beta_t = \varepsilon_t^n$, where $n = 1, 2$ or 3. Obtain the ensemble soft sensor model according to β_t:

$$F_k(x) = \frac{\sum_t (\lg(1/\beta_t)) f_t(x)}{\sum_t (\lg(1/\beta_t))}$$

(7) Calculate the error of F_k using S_k : $E_k = \sum_{ARE_k(i)>\delta} D_t(i)$. If $E_k > \varphi$, give up f_t, and return to step (2).
(8) Calculate $B_k = E_k/(1 - E_k)$ to update the weights:

$$w_{t+1} = w_t \times \begin{cases} B_k & ARE_k(i) < \delta \\ 1 & \text{otherwise} \end{cases}.$$

Output: Obtain the ensemble soft sensor model according to B_k:

$$F_{fin}(x) = \frac{\sum_k (\lg(1/B_k)) F_k(x)}{\sum_k (\lg(1/B_k))}$$

Algorithm 1: Incremental learning ensemble strategy.

After K hypotheses are generated for each subdataset, the final hypothesis F_{fin} is obtained by the weighted majority voting of all the composite hypotheses.

$$F_{fin}(x) = \frac{\sum_k (\lg(1/B_k)) F_k(x)}{\sum_k (\lg(1/B_k))} \tag{7}$$

When new data come constantly during the industrial process, new subdatasets will be generated (they will be the $K + 1$, $K + 2 \ldots$). Based on a new subdataset, a new hypothesized soft sensor can be trained by a new iteration. The new information in the new data will be obtained and added to the final ensemble soft sensor according to (7). As the added incremental learning strategy, the ensemble soft sensor is updated based on the old hypothesis. Therefore, the information in the old data is also retained, and the increment of information from new data is achieved.

Overall, in the above ILES, the ensemble strategy is efficient to improve the prediction accuracy using the changed distribution. Therefore, the ILES will give more attention to the "difficult" data with big errors in every iteration that are attributable to the new distribution. Due to the harder learning for the "difficult" data, more information can be obtained. Therefore, the soft sensor model is built more completely, and the accuracy of perdition is improved. Moreover, the data that is used to train the sublearning machines is divided into a training subdataset and a testing subdataset. The testing error will be used in the following steps: the weight update and the composite hypothesis ensemble. Therefore, the generalization of the soft sensor model based on the ILES can be improved efficiently, especially compared with traditional algorithms. Additionally, when the new data is added, the new ILES with incremental learning ability can learn the new data in real-time and does not give up the old information from the old data. The ILES can save the information of old sublearning machines that have been trained, but it does not need to save the original data. In other words, only a small amount of new production data being saved is enough. This strategy is efficient to save space. Furthermore, the new ILES also may save time compared with the traditional updating method. This strategy is attributed to the conservation of the old B_k and the sublearning machines in composite hypotheses (7).

3. Experiments

In this chapter, the proposed ILES is tested in the sizing production process for predicting the sizing percentage. First, the influence of each parameter on the performance of the proposed algorithm is discussed. Meanwhile, the real industrial process data is used to establish the soft sensor model to verify the incremental learning performance of the algorithm. Finally, to prove its generalization performance, three test functions are used to verify the improvement of the prediction performance. The methods are implemented using MATLAB language and all the experiments are performed on a PC with the Intel Core 7500U CPU (2.70GHZ for each single core) and the Windows 10 operation system.

3.1. Sizing Production Process and Sizing Percentage. The double-dip and double pressure sizing processes are widely used in textile mills, as shown in Figure 1. The sizing percentage plays an important role during the process of sizing for good sizing quality. In addition, the sizing agent control of warp sizing is essential for improving both productivity and product quality. The sizing percentage online detection is a key factor for successful sizing control during the sizing process. The traditional sizing detection methods, which are instruments measurement and indirect calculation, have expensive prices or unsatisfactory accuracy. Soft sensors provide an effective way to predict the sizing percentage and to overcome the above shortages. According to the mechanism analysis of the sizing process, the influencing factors on the sizing percentage are slurry concentration, slurry viscosity, slurry temperature, the pressure of the first Grouting roller, the pressure of the second Grouting roller, the position of immersion roller, the speed of the sizing machine, the cover coefficient of the yarn, the yarn tension, and the drying temperature [22]. In the following soft sensor modeling process, the inputs of soft sensors are the nine influencing factors, and the output is the sizing percentage.

3.2. Experiments for the Parameters of the ILES. Here, we select ELM as the sublearning machine of the ILES, due to its good performance, such as fast learning speed and simple parameter choices [22, 23]; the appendix reviews the process of ELM. Then, experiments with different parameters of the ILES are done to research the performance trend of the ILES when the parameters change.

First, the experiments to assess the ILES algorithm's performance are done with different T_ks. Here, the T_k increases from 1 to 15. Figure 2 shows the results of the training errors and the testing errors with different T_ks. Along with the increasing T_k, the training and testing errors decrease. When T_k increases to 7, the testing error is the smallest. However, when T_k increases to more than 9, the testing error becomes larger again. Furthermore, the training errors only slightly decrease. Therefore, we can draw the conclusion when the parameter T_k is 7 that the performance of ILES is the best. The comparison is also done between AdaBoost.R and the ILES regarding the testing errors with different numbers of ELMs in Figure 3. Although the RMSE means of AdaBoost.R and the ILES are different, their performance trends are similar with the increasing number of ELMs. Here, the RMSE is described as

$$\text{RMSE} = \left(\frac{1}{n} \sum_{i=1}^{n} e^2 (i) \right)^{1/2} \tag{8}$$

Second, we discuss the impact of parameter (δ and φ) changes on the ILES performance. The experiments demonstrate that when δ is too small, the performance of ELM is difficult to achieve the preset goal, and the iteration is difficult to stop. δ is also not larger than 80 percent of the average of the relative errors of ELMs; otherwise the F_k can not be obtained. Furthermore, the value of φ determines the number of "better samples". Here the "better samples" refer to the samples that can reach the expected precision standard of predicted results by submachines. If φ is too small, the ELM soft sensor model (f_t) will not be sufficiently obtained. If φ is too large, the "bad" ELM model (f_t) will be aggregated into the final composite hypothesized $F_{fin}(x)$. Then, the accuracy of the ILES cannot be improved. The relationships among δ, φ, and RMSE are shown in Table 1. When $\delta = 0.06$ and $\varphi = 0.15$, the model has the best performance (the RMSE is 0.3084).

Table 1: The RMSE of the ILES with different parameters Δ, Φ ($T_k = 7$).

δ	φ				
	0.05	0.1	0.15	0.2	0.25
0.02	—	—	—	0.4559	0.4712
0.03	—	—	0.4505	0.4637	0.4614
0.04	—	0.3484	0.3829	0.3888	0.4125
0.05	0.3947	0.3620	0.3765	0.3898	0.3812
0.06	0.4363	0.3734	**0.3084**	0.4036	0.4121
0.07	0.4591	0.3342	0.3323	0.3909	0.3954
0.08	0.4573	0.3682	0.3517	0.4138	0.4332

3.3. Experiments for the Learning Process of the ILES. For establishing the soft sensor model based on the ILES, a total of 550 observations of real production data are collected from Tianjin Textile Engineering Institute Co., Ltd., of which 50 data are selected randomly as testing data. The remaining 500 data are divided into two data sets according to the time of production. The former 450 data are used as the training data set, and the latter 50 data are used as the update data set. The inputs are 9 factors that affect the sizing percentage. The output is the sizing percentage. The parameters of the ILES are $K = 9$, $T_k = 7$, $\delta = 0.06$, and $\varphi = 0.15$. That is to say, the 450 training data are divided into 9 subdatasets $K_1 \sim K_9$, and the number of ELMs is 7. According to the needs of the sizing production, the predictive accuracy of the soft sensors is defined as

$$\text{accuracy} = \frac{N_s}{N_w} \tag{9}$$

where N_s is the number of times with an error < 0.6 and N_w is the total number of testing times.

Since the learning process is similar to the OS-ELM [24] update process. It is an online assessment model that is capable of updating network parameters based on new arriving data without retrains historical data. Therefore, while comparing the accuracy of IELS learning process, it is also compared with OS-ELM. The two columns on the right side of Table 2 show the changes in the soft sensor accuracy during the learning process of the ILES and OS-ELM. It can be seen that the stability and accuracy of ILES are superior to OS-ELM.

3.4. Comparison. In this experiment, we used 10-fold cross validation to test the model's performance. The first 500 data sets are randomly divided into 10 subdatasets $S_1 \sim S_{10}$. The remaining 50 data sets are used as the updated data set S_{11}. The single subdataset from $S_1 \sim S_{10}$ will be retained as the

Table 2: The changes in the soft sensor accuracy during the learning process of the ILES.

Dataset	Iteration										
	1	2	3	4	5	6	7	8	9	Update	OS-ELM
K_1	95%	95%	95%	96%	95%	95%	97%	95%	95%	95%	56%
K_2		94%	94%	96%	94%	95%	95%	94%	94%	94%	68%
K_3			96%	95%	96%	96%	95%	95%	96%	97%	64%
K_4				97%	96%	94%	95%	95%	96%	95%	62%
K_5					96%	94%	96%	95%	95%	96%	66%
K_6						95%	95%	97%	96%	96%	70%
K_7							96%	96%	95%	96%	80%
K_8								93%	94%	94%	88%
K_9									94%	95%	88%
Update set										96%	88%
Testing set	83.2%	85%	85.1%	87%	90%	91%	91.8%	92%	93.2%	95%	90%

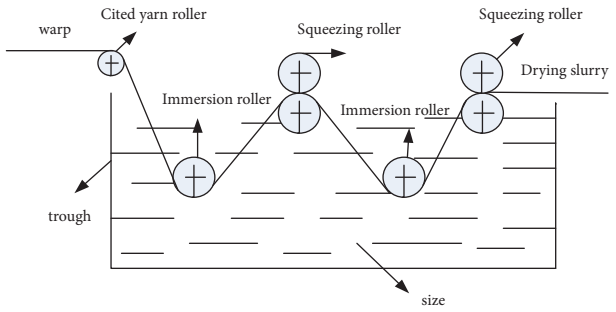


Figure 1: The double-dip and double pressure sizing processes.

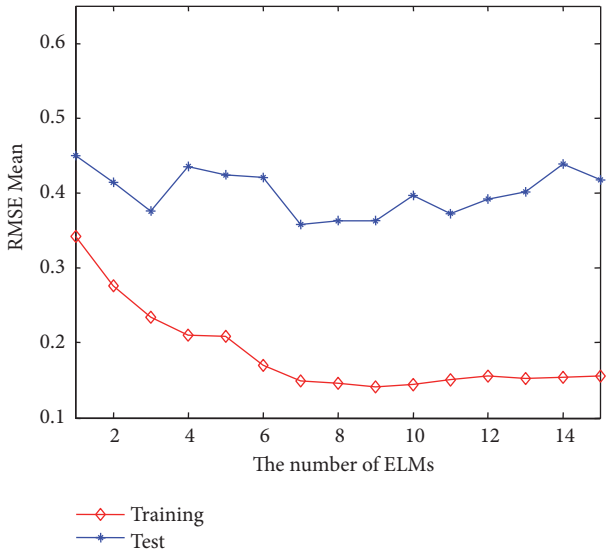


Figure 2: The training and testing errors of the ILES with different parameters T_k.

validation data for testing the model, and the remaining 9 subdatasets are used as the training data. For comparing the new method with other soft sensor methods, the single ELM and ensemble ELM based on AdaBoost.R are also applied to build the sizing percentage soft sensor models as traditional

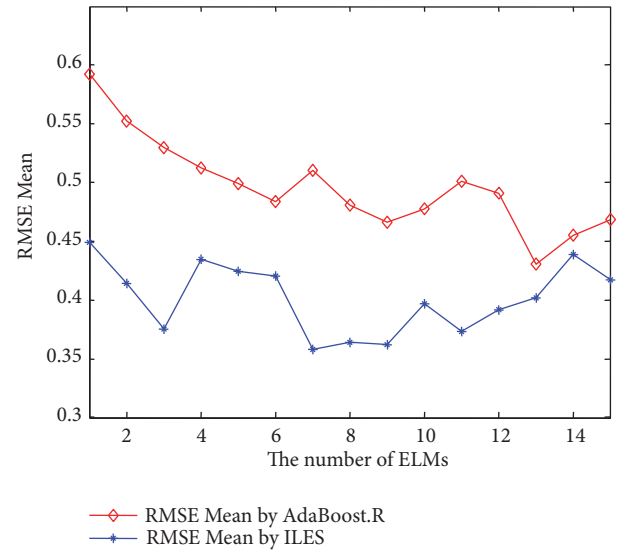


Figure 3: The performance trends of the ILES and AdaBoost.R with different parameters T_k.

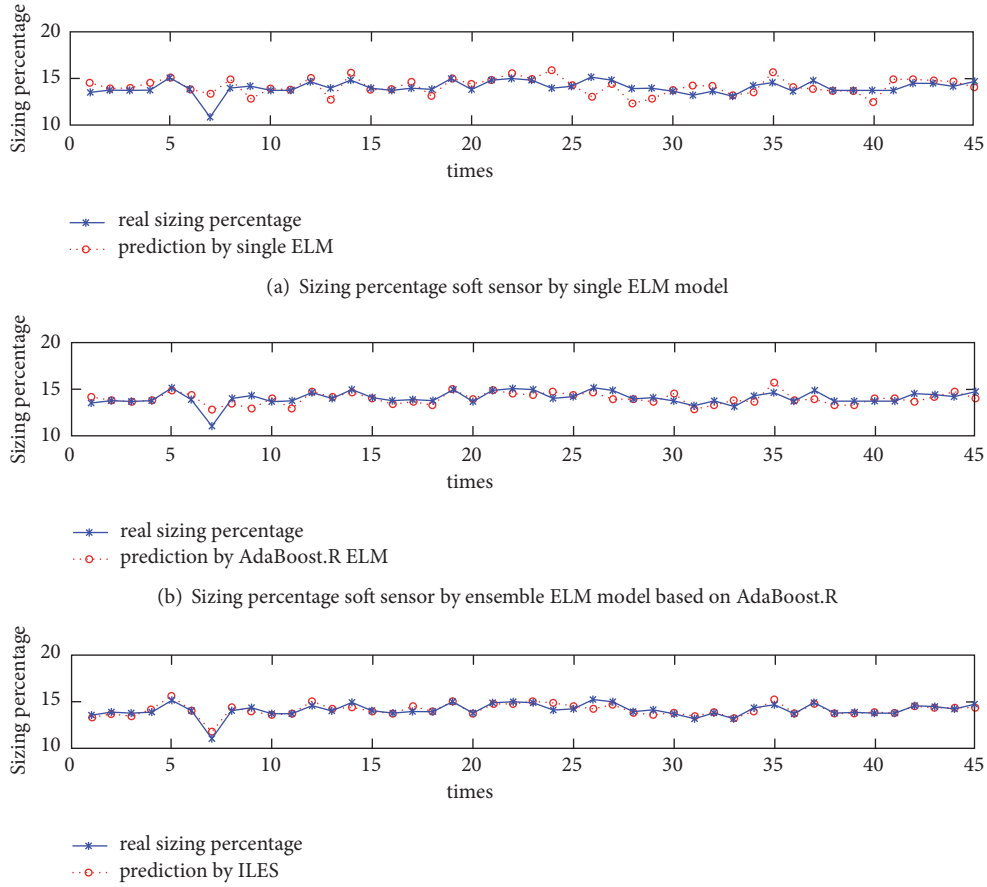
methods with the same data set. The soft sensor models are listed as follows.

Single ELM model: the main factors that affect the sizing percentage are the inputs of the model. The input layer of the ELM has 9 nodes. The hidden layer of the ELM has 2300 nodes, and the output layer of the ELM has one node, which is the sizing percentage.

AdaBoost.R model: the parameters and the structure of the base ELM are the same as those of the single ELM. AdaBoost.R has 13 iterations.

ILES model: the ELMs are same as the single ELM model described above. The parameters of the ILES are $T_k = 7$, $\delta = 0.06$, and $\varphi = 0.15$ (time consuming 163s).

Figures 4(a)–4(c) shows the predicted sizing percentage of different soft sensor models based on different methods. The experiments demonstrate that the strategy of the ILES can improve the accuracy of the soft sensor model. In addition, the training errors of the above three models all



(a) Sizing percentage soft sensor by single ELM model

(b) Sizing percentage soft sensor by ensemble ELM model based on AdaBoost.R

(c) Sizing percentage soft sensor by ILES model

Figure 4: The result of the sizing percentage soft sensor.

can achieve 0.2. However, the testing errors of the prediction models of AdaBoost.R and the ILES are smaller than that of the single ELM. This result means that the ensemble methods have better generalization performance. Table 3 shows the performance of the prediction model based on different methods after updating. The results of the comparison experiments show that the soft sensor based on the new ILES has the best accuracy and the smallest RMSE. This result is attributed to the use of the testing subdataset in the ensemble strategy and the incremental learning strategy during the learning process of the ILES algorithm. Overall, the accuracy of the soft sensor can fit the needs of actual production processes. Moreover, the incremental learning performance can ensure the application of industrial process soft sensors in practical production.

3.5. Experiments for the Performance of the ILES by Test Functions. To verify the universal use of the algorithm, three test functions are used to test the improvement of the prediction performance. These test functions are Friedman#1, Friedman#2, and Friedman#3. Table 4 shows the expression of each test model and the value range of each variable. Friedman#1 has a total of 10 input variables, of which there are five input variables associated with the output variable, and the other five input variables are independent of the output variables. The Friedman#2 and Friedman#3 test functions incorporate the impedance phase change of the AC circuit.

Through continuous debugging, the parameters of each algorithm are determined as shown in Table 5. For every test function, generate a total of 900 data, and 78% of the total samples were selected as training samples, 11% as updating samples and 11% as testing samples, according to the need for different test models. That is to say, the 700 training data are divided into 7 subdatasets $K_1 \sim K_7$. Figures 5–7 show the predicted results of Friedman#1, Friedman#2, and Friedman #3 with different soft sensor models based on different methods (time consuming 227s). The comparison of the performances of the different soft sensors is shown in Table 6. It shows the soft sensor model based on ILES has the best performance.

4. Conclusions

An ILES algorithm is proposed for better accuracy and incremental learning ability for industrial process soft sensors. The sizing percentage soft sensor model is established to test the performance of the ILES. The main factors that influence the sizing percentage are the inputs of the soft sensor model.

Table 3: The performance of the soft sensor model based on different methods with 10-fold cross validation.

Method	datasets			RMSE	ARE	Max ARE
	training	updating	testing			
ELM				0.7987	0.0698	0.2265
AdaBoost.R	$S_1 \sim S_9$	S_{11}	S_{10}	0.5915	0.0313	0.1447
ILES				0.3704	0.0213	0.0689
ELM				0.7386	0.0868	0.3773
AdaBoost.R	$S_1 \sim S_8, S_{10}$	S_{11}	S_9	0.4740	0.0394	0.1846
ILES				0.3309	0.0182	0.0542
ELM				0.8405	0.0772	0.2962
AdaBoost.R	$S_1 \sim S_7, S_9 \sim S_{10}$	S_{11}	S_8	0.5099	0.0295	0.1873
ILES				0.3323	0.0202	0.0572
ELM				0.936	0.0655	0.4255
AdaBoost.R	$S_1 \sim S_6, S_8 \sim S_{10}$	S_{11}	S_7	0.4835	0.219	0.1773
ILES				0.3556	0.0204	0.0525
ELM				0.7342	0.0618	0.3451
AdaBoost.R	$S_1 \sim S_5, S_7 \sim S_{10}$	S_{11}	S_6	0.5198	0.0324	0.1816
ILES				0.3965	0.0221	0.0641
ELM				0.8533	0.0872	0.3546
AdaBoost.R	$S_1 \sim S_4, S_6 \sim S_{10}$	S_{11}	S_5	0.4672	0.0336	0.1724
ILES				0.3531	0.0195	0.0499
ELM				0.7752	0.0749	0.3249
AdaBoost.R	$S_1 \sim S_3, S_5 \sim S_{10}$	S_{11}	S_4	0.4105	0.0362	0.1924
ILES				0.3934	0.0209	0.0748
ELM				0.8430	0.0823	0.3281
AdaBoost.R	$S_1 \sim S_2, S_4 \sim S_{10}$	S_{11}	S_3	0.5773	0.0525	0.1827
ILES				0.3656	0.0203	0.0610
ELM				0.9127	0.1104	0.4302
AdaBoost.R	$S_1, S_3 \sim S_{10}$	S_{11}	S_2	0.5695	0.0580	0.1891
ILES				0.3625	0.0211	0.0545
ELM				0.7905	0.0910	0.2245
AdaBoost.R	$S_2 \sim S_{10}$	S_{11}	S_1	0.5045	0.0436	0.1773
ILES				0.3145	0.0191	0.0756

Table 4: Test function expressions.

Data set	Expression	Variable scope
Friedman#1	$y = 10\sin(\pi x_1 x_2) + 20\left(x_3 - 0.5\right)^2 + 10x_4 + 5x_5 + \delta$	$x_i \sim U[0,1]$, $i = 1, 2..., 10$ $\delta \sim U[-4, 4]$
Friedman#2	$y = \sqrt{{x_1}^2 + \left(x_2 x_3 - \frac{1}{x_2 x_4}\right)^2} + \delta$	$x_1 \sim U[0, 100]$ $x_2 \sim U[40\pi, 560\pi]$ $x_3 \sim U[0, 1]$ $x_4 \sim U[1, 11]$ $\delta \sim U[-120, 120]$
Friedman#3	$y = \tan^{-1}\left(\frac{x_2 x_3 - 1/x_2 x_4}{x_1}\right) + \delta$	$x_1 \sim U[0, 100]$ $x_2 \sim U[40\pi, 560\pi]$ $x_3 \sim U[0, 1]$ $x_4 \sim U[1, 11]$ $\delta \sim U[-0.4, 0.4]$

Table 5: Parameters of the algorithmic performance test.

Test function	T_k (the number of ELMs)	K	δ	φ	datasets		
					training	updating	testing
Friedman#1	10	7	0.012	0.1050	700	100	100
Friedman#2	10	7	0.012	0.510	700	100	100
Friedman#3	8	7	0.010	0.1810	700	100	100

Table 6: The performance comparisons of the algorithms.

Data set	Method	RMSE	ARE	Max ARE
Friedman#1	ELM	0.8278	0.7662	4.7880
	AdaBoost.R	0.7987	0.6536	3.8817
	ILES	0.3099	0.3073	2.7132
Friedman#2	ELM	0.6338	0.9975	17.9870
	AdaBoost.R	0.4190	0.4495	7.3544
	ILES	0.2079	0.3683	7.2275
Friedman#3	ELM	1.1409	1.4314	14.9348
	AdaBoost.R	0.8959	1.3461	11.5991
	ILES	0.4624	0.3930	5.2719

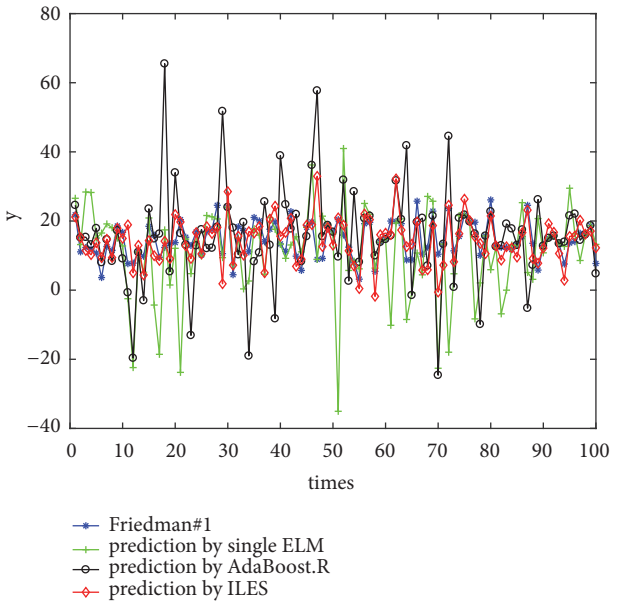


Figure 5: The comparison of Friedman#1 with different models.

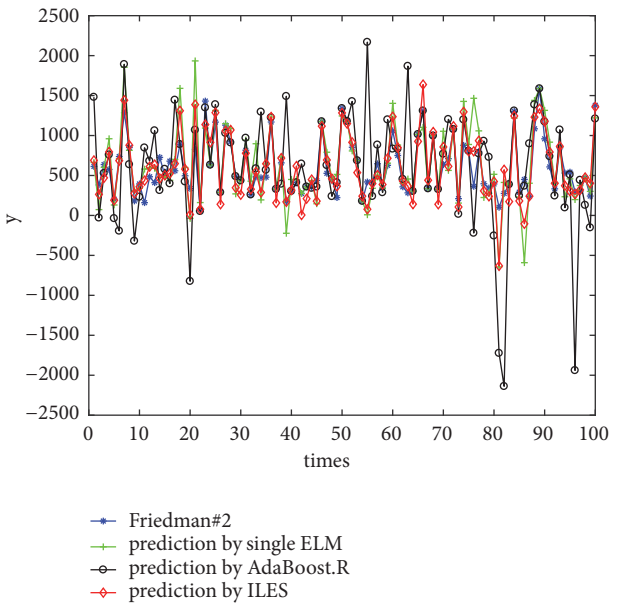


Figure 6: The comparison of Friedman#2 with different models.

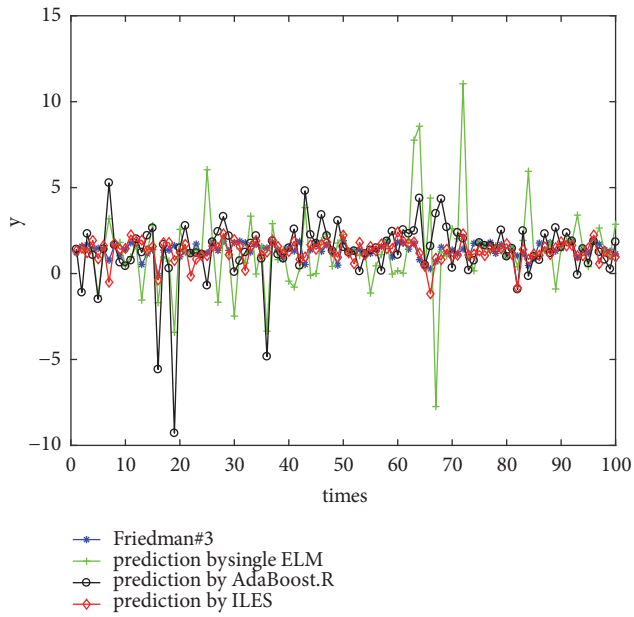


Figure 7: The comparison of Friedman#3 with different models.

Then, the ensemble model is trained with different subtraining dataset, and a soft sensor model with incremental learning performance is obtained by the ILES strategy. When new data add up to a certain number, the model will be updated using an incremental learning strategy. The new sizing percentage soft sensor model is used in Tianjin Textile Engineering Institute Co., Ltd. The experiments demonstrate that the new soft sensor model based on the ILES has good performance. The predictive accuracy of the new soft sensor model could be satisfied for sizing production. Finally, the new ILES is also tested with three testing functions to verify the performance with different data sets for universal use. Because the size of the subdataset is different from the experiment on sizing percentage, it can be conclude from the prediction results that the size of subdataset does not affect the performance of the algorithm. In the future, the ILES can also be used in other complex industry processes that require the use of soft sensors.

Appendix

Review of Extreme Learning Machine

Single Hidden Layer Feedforward Networks (SLFNs) with Random Hidden Nodes

For N arbitrary distinct samples $\{(x_j, t_j)\}_{j=1}^{N}$, where $x_j = [x_{j1}, x_{j2}, \cdots x_{jn}]^T \in R^n$ and $t_j = [t_{j1}, t_{j2}, \cdots t_{jm}]^T \in R^n$, standard SLFNs with $\widetilde{N}$ hidden nodes and the activation function $g(x)$ are mathematically modeled as

$$\sum_{i=1}^{\widetilde{N}} \beta_i g_i\left(x_j\right) = \sum_{i=1}^{\widetilde{N}} \beta_i g\left(w_i \cdot x_j + b_i\right) = o_j, \quad j = 1, \cdots, N \tag{A.1}$$

where $w_i = [w_{i1}, w_{i2}, \cdots w_{in}]^T$ is the weight vector connecting the ith hidden node and the input nodes, $\beta_i = [\beta_{i1}, \beta_{i2}, \cdots \beta_{im}]^T$ is the weight vector connecting the ith hidden node and the output nodes, $o_j = [o_{j1}, o_{j2}, \cdots, o_{jm}]^T$ is the output vector of the SLFN, and b_i is the threshold of the ith hidden node. $w_i \cdot x_j$ denotes the inner product of w_i and x_j. The output nodes are chosen linearly. The standard SLFNs with $\widetilde{N}$ hidden nodes with the activation function $g(x)$ can approximate these N samples with zero error means such that $\sum_{j=1}^{\widetilde{N}} \|o_j - t_j\| = 0$. These N equations can be written compactly as follows:

$$\mathbf{H}\beta = \mathbf{T} \tag{A.2}$$

where

$$\mathbf{H} = \begin{bmatrix} g\left(w_1 \cdot x_1 + b_1\right) & \cdots & g\left(w_{\widetilde{N}} \cdot x_1 + b_{\widetilde{N}}\right) \\ \vdots & \cdots & \vdots \\ g\left(w_1 \cdot x_N + b_1\right) & \cdots & g\left(w_{\widetilde{N}} \cdot x_N + b_{\widetilde{N}}\right) \end{bmatrix}_{N \times \widetilde{N}} \tag{A.3}$$

$$\beta = \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_{\widetilde{N}}^T \end{bmatrix}_{\widetilde{N} \times m} \tag{A.4}$$

and

$$\mathbf{T} = \begin{bmatrix} t_1^T \\ \vdots \\ t_N^T \end{bmatrix}_{N \times m} \tag{A.5}$$

Here, $\mathbf{H}$ is the hidden layer output matrix.

ELM Algorithm. The parameters of the hidden nodes do not need to be tuned and can be randomly generated permanently according to any continuous probability distribution. The unique smallest norm least squares solution of the above linear system is

$$\widehat{\beta} = \mathbf{H}^{\dagger}\mathbf{T} \tag{A.6}$$

where $\mathbf{H}^{+}$ is the Moore-Penrose generalized inverse of matrix $\mathbf{H}$.

Thus, a simple learning method for SLFNs, called extreme learning machine (ELM), can be summarized as follows.

Step 1. Randomly assign input weight w_i and bias b_i, $i = 1, \cdots, \widetilde{N}$,

Step 2. Calculate the hidden layer output matrix $\mathbf{H}$.

Step 3. Calculate the output weight $\beta : \beta = \mathbf{H}^{\dagger}\mathbf{T}$.

The universal approximation capability of the ELM has been rigorously proved in an incremental method by Huang *et al*. [23].

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research is supported by the National Natural Science Foundation of China (Grants nos. 71602143, 61403277, 61573086, and 51607122), Tianjin Natural Science Foundation (no. 18JCYBJC22000), Tianjin Science and Technology Correspondent Project (no. 18JCTPJC62600), the Program for Innovative Research Team in University of Tianjin (nos. TD13-5038, TD13-5036), and State Key Laboratory of Process Automation in Mining & Metallurgy/Beijing Key Laboratory of Process Automation in Mining & Metallurgy Research Fund Project (BGRIMM-KZSKL-2017-01).

References

[1] R. E. Schapire, "The strength of weak learnability," *Machine Learning*, vol. 5, no. 2, pp. 197–227, 1990.

[2] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[3] S. Liu, S. Wang, J. Chen, X. Liu, and H. Zhou, "Moving larval shrimps recognition based on improved principal component analysis and AdaBoost," *Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE)*, vol. 33, no. 1, pp. 212–218, 2017.

[4] Y. Li, H. Guo, and X. Liu, "Classification of an integrated algorithm based on Boosting in unbalanced data," *Journal of Systems Engineering and Theory*, vol. 01, pp. 189–199, 2016.

[5] L. L. Wang and Z. L. Fu, "AdaBoost algorithm for multi-tag classification based on label correlation," *Journal of Sichuan University (Engineering Science Edition)*, vol. 5, pp. 91–97, 2016.

[6] H. Tian and A. Wang, "Advanced AdaBoost modeling method based on incremental learning," *Control and Decision*, vol. 09, pp. 1433–1436, 2012.

[7] M. Wu, J. Guo, Y. Ju, Z. Lin, and Q. Zou, "Parallel algorithm for parallel selection based on hierarchical filtering and dynamic updating," *Computer Science*, vol. 44, no. 1, pp. 48–52, 2017.

[8] H. Drucker, "Improving regressor using boosting," in *Proceedings of the 14th International Conference on Machine Learning*, pp. 107–115, Morgan Kaufmann Publishers, Burlington, Mass, USA, 1997.

[9] R. Avnimelech and N. Intrator, "Boosting regression estimators," *Neural Computation*, vol. 11, no. 2, pp. 499–520, 1999.

[10] R. Feely, *Predicting Stock Market Volatility Using Neural Networks*, Trinity College Dublin, 2000.

[11] D. L. Shrestha and D. P. Solomatine, "Experiments with AdaBoost.RT, an improved boosting scheme for regression," *Neural Computation*, vol. 18, no. 7, pp. 1678–1710, 2006.

[12] H.-X. Tian and Z.-Z. Mao, "An ensemble ELM based on modified AdaBoost.RT algorithm for predicting the temperature of molten steel in ladle furnace," *IEEE Transactions on Automation Science and Engineering*, vol. 7, no. 1, pp. 73–80, 2010.

[13] L. Kao, C. Chiu, C. Lu, C. Chang et al., "A hybrid approach by integrating wavelet-based feature extraction with MARS and SVR for stock index forecasting," *Decision Support Systems*, vol. 54, no. 3, pp. 1228–1244, 2013.

[14] X. Qiu, P. N. Suganthan, and G. A. J. Amaratunga, "Ensemble incremental learning Random Vector Functional Link network for short-term electric load forecasting," *Knowledge-Based Systems*, vol. 145, no. 4, pp. 182–196, 2018.

[15] Z. Zhou, J. Chen, and Z. Zhu, "Regularization incremental extreme learning machine with random reduced kernel for regression," *Neurocomputing*, vol. 321, no. 12, pp. 72–81, 2018.

[16] R. Polikar, L. Udpa, S. S. Udpa, and V. Honavar, "Learn++: an incremental learning algorithm for supervised neural networks," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 31, no. 4, pp. 497–508, 2001.

[17] G. Ditzler and R. Polikar, "Incremental learning of concept drift from streaming imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 10, pp. 2283–2301, 2013.

[18] D. Brzezinski and J. Stefanowski, "Reacting to different types of concept drift: the accuracy updated ensemble algorithm," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 1, pp. 81–94, 2013.

[19] H.-J. Rong, Y.-X. Jia, and G.-S. Zhao, "Aircraft recognition using modular extreme learning machine," *Neurocomputing*, vol. 128, pp. 166–174, 2014.

[20] Q. Lin, X. Wang, and H.-J. Rong, "Self-organizing fuzzy failure diagnosis of aircraft sensors," *Memetic Computing*, vol. 7, no. 4, pp. 243–254, 2015.

[21] X. Xu, D. Jiang, B. Li, and Q. Chen, "Optimization of Gaussian mixture model motion detection algorithm," *Journal of Chemical Industry and Engineering*, vol. 55, no. 6, pp. 942–946, 2004.

[22] H. Tian and Y. Jia, "Online soft measurement of sizing percentage based on integrated multiple SVR fusion by Bagging," *Journal of Textile Research*, vol. 35, no. 1, pp. 62–66, 2014 (Chinese).

[23] G. Huang, Q. Zhu, and C. Siew, "Extreme learning machine: a new learning scheme of feedforward neural networks," in *Proceedings of International Joint Conference on Neural Networks*, vol. 2, pp. 985–990, 2004.

[24] N. Liang, G. Huang, H. Rong et al., "On-line sequential extreme learning machine," in *Proceedings of the Iasted International Conference on Computational Intelligence*, pp. 232–237, DBLP, Calgary, Alberta, Canada, 2005.

Hindawi
Complexity
Volume 2019, Article ID 8523748, 16 pages
https://doi.org/10.1155/2019/8523748

Research Article

Looking for Accurate Forecasting of Copper TC/RC Benchmark Levels

Francisco J. Díaz-Borrego, María del Mar Miras-Rodríguez, and Bernabé Escobar-Pérez

Universidad de Sevilla, Seville, Spain

Correspondence should be addressed to María del Mar Miras-Rodríguez; mmiras@us.es

Received 22 November 2018; Revised 24 February 2019; Accepted 5 March 2019; Published 1 April 2019

Guest Editor: Marisol B. Correia

Forecasting copper prices has been the objective of numerous investigations. However, there is a lack of research about the price at which mines sell copper concentrate to smelters. The market reality is more complex since smelters obtain the copper that they sell from the concentrate that mines produce by processing the ore which they have extracted. It therefore becomes necessary to thoroughly analyse the price at which smelters buy the concentrates from the mines, besides the price at which they sell the copper. In practice, this cost is set by applying discounts to the price of cathodic copper, the most relevant being those corresponding to the smelters' benefit margin (*Treatment Charges-TC* and *Refining Charges-RC*). These discounts are agreed upon annually in the markets and their correct forecasting will enable making more adequate models to estimate the price of copper concentrates, which would help smelters to duly forecast their benefit margin. Hence, the aim of this paper is to provide an effective tool to forecast copper TC/RC annual benchmark levels. With the annual benchmark data from 2004 to 2017 agreed upon during the LME Copper Week, a three-model comparison is made by contrasting different measures of error. The results obtained indicate that the LES (*Linear Exponential Smoothing*) model is the one that has the best predictive capacity to explain the evolution of TC/RC in both the long and the short term. This suggests a certain dependency on the previous levels of TC/RC, as well as the potential existence of cyclical patterns in them. This model thus allows us to make a more precise estimation of copper TC/RC levels, which makes it useful for smelters and mining companies.

1. Introduction

1.1. Background. The relevance of copper trading is undeniable. In 2016 exports of copper ores, concentrates, copper matte, and cement copper increased by 1.5%, reaching 47.3 b $USD, while imports attained 43.9 b $USD [1]. In addition, the global mining capacity is expected to rise by 10% from the 23.5 million tonnes recorded in 2016 to 25.9 million tonnes in 2020, with smelter production having reached the record figure of 19.0 million tonnes in 2016 [2].

The world's copper production is essentially achieved through alternative processes which depend on the chemical and physical characteristics of the copper ores extracted. According to the USGS' 2017 Mineral Commodity Summary on Copper [3], global identified copper resources contained 2.1 billion tonnes of copper as of 2014, of which about 80% are mainly copper sulphides, whose copper content has to be extracted through pyrometallurgical processes [4]. In 2010, the average grades of ores being mined ranged from 0.5% to 2% Cu, which makes direct smelting unfeasible for economic and technical reasons. So, copper sulphide ores undergo a process known as froth-flotation to obtain concentrates containing ≈ 30% Cu, which makes concentrates the main products offered by copper mines [2, 5]. Concentrates are later smelted and, in most cases, electrochemically refined to produce high-purity copper cathodes (Figure 1). Copper cathodes are generally regarded as pure copper, with a minimum copper content of 99.9935% Cu [6]. Cathodes are normally produced by integrated smelters that purchase concentrates at a discounted price of the copper market price and sell cathodic copper at the market price, adding a premium when possible.

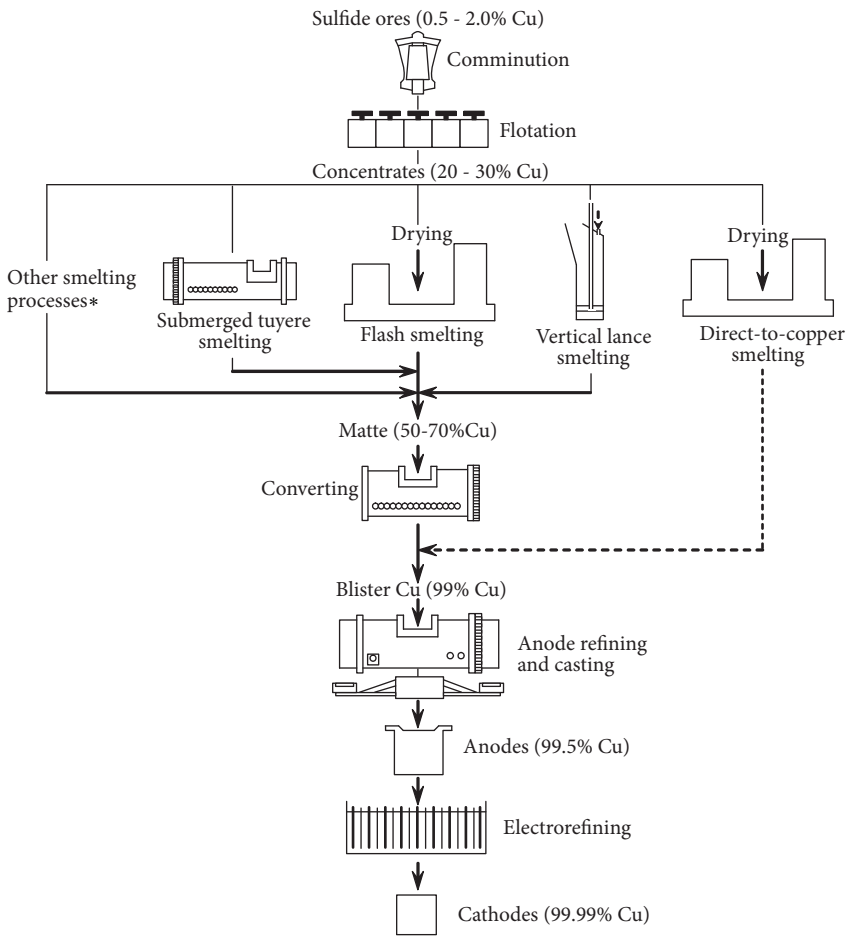


Figure 1: Industrial processing of copper sulphides ore to obtain copper cathodes. (Source: Extractive Metallurgic of Copper, pp.2 [4]).

1.2. TC/RC and Their Role in Copper Concentrates Trade. The valuation of these copper concentrates is a recurrent task undertaken by miners or traders following processes in which market prices for copper and other valuable metals such as gold and silver are involved, as well as relevant discounts or coefficients that usually represent a major part of the revenue obtained for concentrate trading, smelting, or refining. The main deductions are applied to the market value of the metal contained by concentrates such as *Copper Treatment Charges (TC)*, *Copper Refining Charges (RC)*, *the Price Participation Clause (PP)*, and *Penalties for Punishable Elements* [7]. These are fixed by the different parties involved in a copper concentrates long-term or spot supply contract, where TC/RC are fixed when the concentrates are sold to a copper smelter/refinery. The sum of TC/RC is often viewed as the main source of revenue for copper smelters along with copper premiums linked to the selling of copper cathodes. Furthermore, TC/RC deductions pose a concern for copper mines as well as a potential arbitrage opportunity for traders, whose strong financial capacity and indepth knowledge of the market make them a major player [8].

Due to their nature, TC/RC are discounts normally agreed upon taking a reference which is traditionally set on an annual basis at the negotiations conducted by the major market participants during *LME Week* every October and, more recently, during the *Asia Copper Week* each November, an event that is focused more on Chinese smelters. The TC/RC levels set at these events are usually taken as benchmarks for the negotiations of copper concentrate supply contracts throughout the following year. Thus, as the year goes on, TC/RC average levels move up and down depending on supply and demand, as well as on concentrate availability and smelters' available capacity. Consultants, such as *Platts, Wood Mackenzie*, and *Metal Bulletin,* regularly carry out their own market surveys to estimate the current TC/RC levels. Furthermore, *Metal Bulletin* has created the first TC/RC index for copper concentrates [9].

1.3. The Need for Accurate Predictions of TC/RC. The current information available for market participants may be regarded as sufficient to draw an accurate assumption of market sentiment about current TC/RC levels, but not enough to foresee potential market trends regarding these crucial discounts, far less as a reliable tool which may be ultimately applicable by market participants to their decision-making framework or their risk-management strategies. Hence, from an organisational standpoint, providing accurate forecasts of copper TC/RC benchmark levels, as well as an accurate

mathematical model to render these forecasts, is a research topic that is yet to be explored in depth. This is a question with undeniable economic implications for traders, miners, and smelters alike, due to the role of TC/RC in the copper trading revenue stream.

Our study seeks to determine an appropriate forecasting technique for TC/RC benchmark levels for copper concentrates that meets the need of reliability and accuracy. To perform this research, three different and frequently-applied techniques have been preselected from among the options available in the literature. Then, their forecasting accuracy at different time horizons will be tested and compared. These techniques (Geometric Brownian Motion -GBM-; the Mean Reversion -MR-; Linear Exponential Smoothing -LES-), have been chosen primarily because they are common in modelling commodities prices and their future expected behaviour, as well as in stock indices' predictive works, among other practical applications [10–13]. The selection of these models is further justified by the similarities shared by TC/RC with indices, interest rates, or some economic variables that these models have already been applied to. Also in our view, the predictive ability of these models in commodity prices such as copper is a major asset to take them into consideration. The models have been simulated using historical data of TC/RC annual benchmark levels from 2004 to 2017 agreed upon during the LME Copper Week. The dataset employed has been split into two parts, with two-thirds as the in-sample dataset and one-third as the out-of-sample one.

The main contribution of this paper is to provide a useful and applicable tool to all parties involved in the copper trading business to forecast potential levels of critical discounts to be applied to the copper concentrates valuation process. To carry out our research, we have based ourselves on the following premises: (1) GBM would deliver good forecasts if copper TC/RC benchmark levels vary randomly over the years, (2) a mean-reverting model, such as the OUP, would deliver the best forecasts if TC/RC levels were affected by market factors and consequently they move around a long-term trend, and (3) a moving average model would give a better forecast than the other two models if there were a predominant factor related to precedent values affecting the futures ones of benchmark TC/RC. In addition, we have also studied the possibility that a combination of the models could deliver the most accurate forecast as the time horizon considered is increased, since there might thus be a limited effect of past values, or of sudden shocks, on future levels of benchmark TC/RC. So, after some time, TC/RC levels could be "normalized" towards a long-term average.

The remainder of this article is structured as follows: Section 2 revises the related work on commodity discounts forecasting and commodity prices forecasting techniques, as well as different forecasting methods; Section 3 presents the reasoning behind the choice of each of the models employed, as well as the historic datasets used to conduct the study and the methodology followed; Section 4 shows the results of simulations of different methods; Section 5 indicates error comparisons to evaluate the best forecasting alternative for TC/RC among all those presented; Section 6 contains the study's conclusions and proposes further lines of research.

2. Related Work

The absence of any specific method in the specialised literature in relation to copper TC/RC leads us to revisit previous literature in order to determine the most appropriate model to employ for our purpose, considering those that have already been used in commodity price forecasting as the logical starting point due to the application similarities that they share with ours.

Commodity prices and their forecasting have been a topic intensively analysed in much research. Hence, there are multiple examples in literature with an application to one or several commodities, such as Xiong et al. [14], where the accuracy of different models was tested to forecast the interval of agricultural commodity future prices; Shao and Dai [15], whose work employs the Autoregressive Integrated Moving Average (ARIMA) methodology to forecast food crop prices such as wheat, rice, and corn; and Heaney [16], who tests the capacity of commodities future prices to forecast their cash price were the cost of carry to be included in considerations, using the LME Lead contract as an example study. As was done in other research [17–19], Table 1 summarizes similar research in the field of commodity price behaviours.

From the summary shown in Table 1, it is seen that moving average methods have become increasingly popular in commodity price forecasting. The most broadly implemented moving average techniques are ARIMA and Exponential Smoothing. They consider the entire time series data and do not assign the same weight to past values than those closer to the present as they are seen as affecting greater to future values. The Exponential Smoothing models' predictive accuracy has been tested [20, 21], concluding that there are small differences between them (Exponential Smoothing) and ARIMA models.

Also, GBM and MR models have been intensively applied to commodity price forecasting. Nonetheless, MR models present a significant advantage over GBM models which allows them to consider the underlying price trend. This advantage is of particular interest for commodities that, according to Dixit and Pindyck [22]—pp. 74, regarding the price of oil "*in the short run, it might fluctuate randomly up and down (in responses to wars or revolutions in oil producing countries, or in response to the strengthening or weakening of the OPEC cartel), in the longer run, it ought to be drawn back towards the marginal cost of producing oil. Thus, one might argue that the price of oil should be modelled as a mean-reverting process.*"

3. Methodology

This section presents both the justification of the models chosen in this research and the core reference dataset used as inputs for each model compared in the methodology, as well as the steps followed during the latter to carry out an effective comparison in terms of the TC/RC forecasting ability of each of these models.

3.1. Models in Methodology. GBM (see *Appendix A* for models definition.) has been used in much earlier research as a

Table 1: Summary of previous literature research.

AUTHOR	RESEARCH
Shafiee & Topal [17]	Validates a modified version of the long-term trend reverting jump and dip diffusion model for forecasting commodity prices and estimates the gold price for the next 10 years using historical monthly data.
Li *et al.* [18]	Proposes an ARIMA-Markov Chain method to accurately forecast mineral commodity prices, testing the method using mineral molybdenum prices.
Issler *et al.* [19]	Investigates several commodities' co-movements, such as Aluminium, Copper, Lead, Nickel, Tin, and Zinc, at different time frequencies, and uses a bias-corrected average forecast method proposed by Issler and Lima [43] to give combined forecasts of these metal commodities employing RMSE as a measure of forecasting accuracy.
Hamid & Shabri [44]	Models palm oil prices using the Autoregressive Distributed Lag (ARDL) model and compares its forecasting accuracy with the benchmark model ARIMA. It uses an ARDL bound-testing approach to co-integration in order to analyse the relationship between the price of palm oil and its determinant factors.
Duan *et al.* [45]	Predicts China's crude oil consumption for 2015-2020 using the fractional-order FSIGM model.
Brennan & Schwartz [46]	Employs the Geometric Brownian Motion (GBM) to analyse a mining project's expected returns assuming it produces a single commodity.
McDonald & Siegel [47]	Uses GBM to model the random evolution of the present value of an undefined asset in an investment decision model.
Zhang *et al.* [48]	Models gold prices using the Ornstein-Uhlenbeck Process (OUP) to account for a potentially existent long-term trend in a Real Option Valuation of a mining project.
Sharma [49]	Forecasts gold prices in India with the Box Jenkins ARIMA method.

way of modelling prices that are believed not to follow any specific rule or pattern and hence seen as *random*. Black and Scholes [23] first used GBM to model stock prices and since then others have used it to model asset prices as well as commodities, these being perhaps the most common of all, in which prices are expected to increase over time, as does their variance [11]. Hence, following our first premise, concerning whether TC/RC might vary randomly, there should not exist a main driving factor that would determine TC/RC future benchmark levels and therefore GBM could to a certain extent be a feasible model for them.

However, although GBM or "random walk" may be well suited to modelling immediate or short-term price paths for commodities, or for TC/RC in our case, it lacks the ability to include the underlying long-term price trend should we assume that there is one. Thus, in accordance with our second premise on benchmark TC/RC behaviour, levels would move in line with copper concentrate supply and demand as well as the smelters' and refineries' available capacity to transform concentrates into metal copper. Hence, a relationship between TC/RC levels and copper supply and demand is known to exist and, therefore, is linked to its market price, so to some extent they move together coherently. Therefore, in that case, as related works on commodity price behaviour such as Foo, Bloch, and Salim [24] do, we have opted for the MR model, particularly the OUP model.

Both GBM and MR are Markov processes which means that future values depend exclusively on the current value, while the remaining previous time series data are not considered. On the other hand, moving average methods employ the average of a pre-established number of past values in different ways, evolving over time, so future values do not rely exclusively on the present, hence behaving as though they had only a limited memory of the past. This trait of the moving average model is particularly interesting when past prices are believed to have a certain, though limited, effect on present values, which is another of the premises for this research. Existing alternatives of moving average techniques pursue considering this "memory" with different approaches. As explained by Kalekar [25], Exponential Smoothing is suitable only for the behaviours of a specific time series; thus Single Exponential Smoothing (SES) is reasonable for short-term forecasting with no specific trend in the observed data, whereas Double Exponential Smoothing or Linear Exponential Smoothing (LES) is appropriate when data shows a cyclical pattern or a trend. In addition, seasonality in observed data can be computed and forecasted through the usage of Exponential Smoothing by the Holt-Winters method, which adds an extra parameter to the model to handle this characteristic.

3.2. TC/RC Benchmark Levels and Sample Dataset. TC/RC levels for copper concentrates continuously vary throughout the year, relying on private and individual agreements between miners, traders, and smelters worldwide. Nonetheless, the TC/RC benchmark fixed during the *LME week* each October is used by market participants as the main reference to set actual TC/RC levels for each supply agreed upon for the following year. Hence, the year's benchmark TC/RC is taken here as a good indicator of a year's TC/RC average levels. Analysed time series of benchmark TC/RC span from 2004 through to 2017, as shown in Table 2, as well as the source each value was obtained from. We have not intended to reflect the continuous variation of TC/RC for the course of any given year, though we have however considered benchmark prices alone as we intuitively assume that annual variations of actual TC/RC in contracts will eventually be reflected in the benchmark level that is set at the end of the year for the year to come.

3.3. TC/RC Relation. TC and RC normally maintain a 10:1 relation with different units, with TC being expressed in US

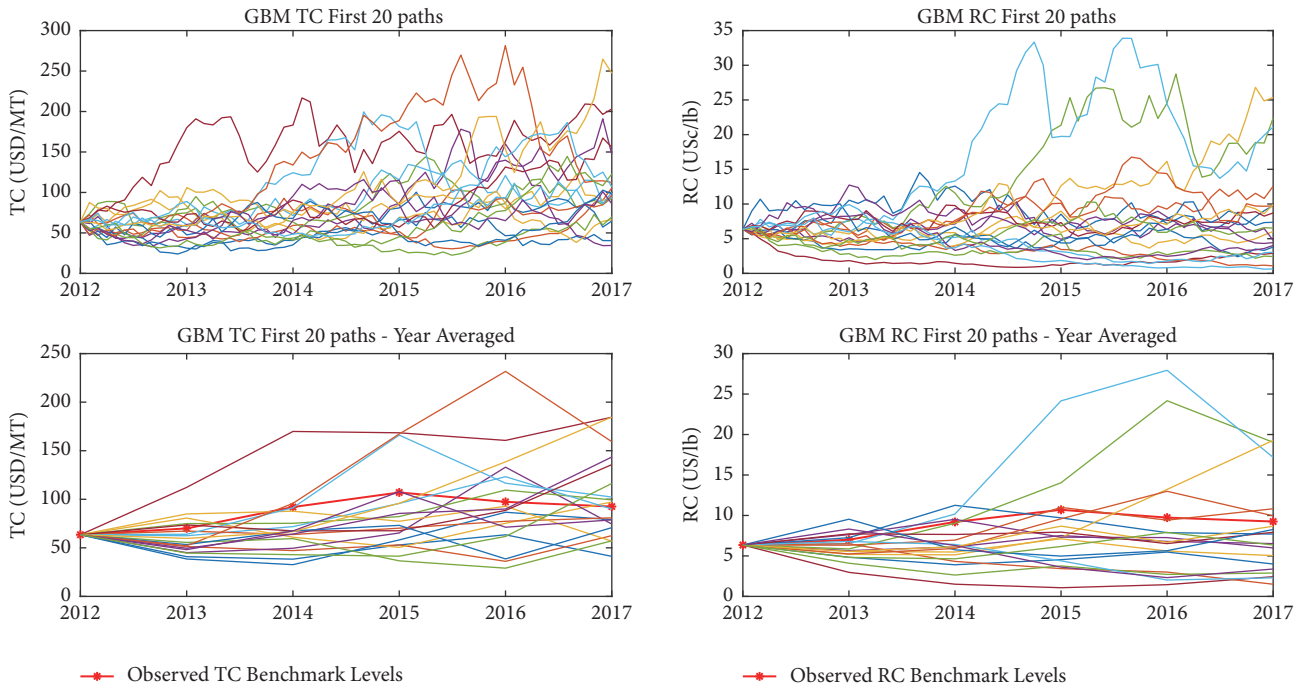


Figure 2: The upper illustrations show the first 20 Monte Carlo paths for either TC or RC levels using monthly steps. The illustrations below show the annual averages of monthly step forecasts for TC and RC levels.

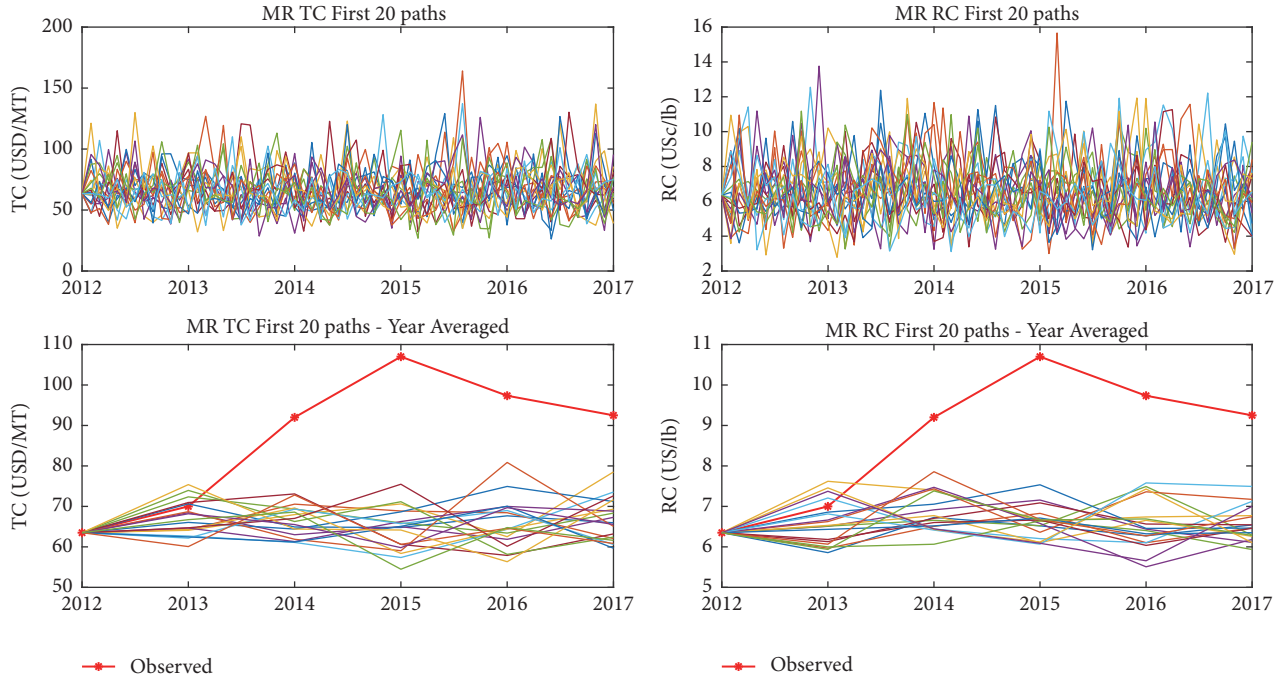


Figure 3: The first 20 Monte Carlo paths following the OUP model using monthly steps are shown on the upper illustrations for either TC or RC. The annual averaged simulated paths every 12 values are shown below.

dollars per metric tonne and RC in US cents per pound of payable copper content in concentrates. In fact, benchmark data show that, of the last 14 years, it was only in 2010 that values of TC and RC did not conform to this relation, though it did remain close to it (46.5/4.7). Nonetheless, this relation has not been observed in the methodology herein, thus treating TC and RC independently, developing separate unrelated forecasts for TC and RC to understand whether

Table 2: TC/RC year benchmark levels (Data available free of charge. Source: Johansson [50], Svante [51], Teck [52], Shaw [53], Willbrandt and Faust [54, 55], Aurubis AG [56], Drouven and Faust [57], Aurubis AG [58], EY [59], Schachler [60], Nakazato [61].).

YEAR	TC (USD/MT)	RC (USc/lb)
2004	45	4.5
2005	85	8.5
2006	95	9.5
2007	60	6.0
2008	45	4.5
2009	75	7.5
2010	46.5	4.7
2011	56	5.6
2012	63.5	6.35
2013	70	7.0
2014	92	9.2
2015	107	10.7
2016	97.35	9.735
2017	92.5	9.25

these individual forecasts for TC and RC would render better results than a single joint forecast for both TC/RC levels that would take into account the 10:1 relation existent in the data.

3.4. Models Comparison Method. The works referred to in the literature usually resort to different measures of errors to conduct the testing of a model's forecasting capacity regardless of the specific nature of the forecasted value, be it cotton prices or macroeconomic parameters. Thus, the most widely errors used include *Mean Squared Error (MSE), Mean Absolute Deviation (MAD), Mean Absolute Percentage Error (MAPE)*, and *Root Mean Square Error (RMSE)* [14, 16, 17, 21, 25–29]. In this work Geometric Brownian Motion (GBM), Ornstein-Uhlenbeck Process (OUP) and Linear Exponential Smoothing (LES) models have been used to forecast annual TC/RC benchmark levels, using all the four main measures of error mentioned above to test the predictive accuracy of these three models. The GBM and OUP models have been simulated and tested with different step sizes, while the LES model has been analysed solely using annual time steps.

The GBM and OUP models were treated separately to the LES model; thus GBM and OUP were simulated using Monte Carlo (MC) simulations, whereas LES forecasts were simple calculations. In a preliminary stage, the models were first calibrated to forecast values from 2013 to 2017 using available data from 2004 to 2012 for TC and RC separately, hence the disregarding of the well-known inherent 10:1 relation (*see Appendix B for Models Calibration*). The GBM and OUP models were calibrated for two different step sizes, monthly steps and annual steps, in order to compare forecasting accuracy with each step size in each model. Following the calibration of the models, Monte Carlo simulations (MC) were carried out using Matlab software to render the pursued forecasts of GBM and OUP models, obtaining 1000 simulated paths for each step size. MC simulations using monthly steps for the 2013–2017 timespan were averaged every 12 steps to deliver year forecasts of TC/RC benchmark levels. On the other hand, forecasts obtained by MC simulations taking annual steps for the same period were considered as year forecasts for TC/RC annual benchmark levels without the need for extra transformation. Besides, LES model forecasts were calculated at different time horizons to be able to compare the short-term and long-term predictive accuracy. LES forecasts were obtained for one-year-ahead, hence using known values of TC/RC from 2004 to 2016; for two years ahead, stretching the calibrating dataset from 2004 to 2015; for five-year-ahead, thus using the same input dataset as for the GBM and OUP models, from 2004 to 2012.

Finally, for every forecast path obtained, we have calculated the average of the squares of the errors with respect to the observed values, *MSE*, the average distance of a forecast to the observed mean, *MAD*, the average deviation of a forecast from observed values, *MAPE*, and the square root of MSE, *RMSE* (see *Appendix C* for error calculations.). The results of *MSE, MAD, MAPE*, and *RMSE* calculated for each forecast path were averaged by the total number of simulations carried out for each case. Averaged values of error measures of all simulated paths, $\overline{\text{MSE}}$, $\overline{\text{MAD}}$, $\overline{\text{MAPE}}$, and $\overline{\text{RMSE}}$, for both annual-step forecasts and monthly step forecasts have been used for cross-comparison between models to test predictive ability at every step size possible.

Also, to test each model's short-term forecasting capacity against its long-term forecasting capacity, one-year-ahead forecast errors of the LES model were compared with the errors from the last year of the GBM and OUP simulated paths. Two-year-ahead forecast errors of LES models were compared with the average errors of the last two years of GBM and OUP, and five-year-ahead forecast errors of LES models were compared with the overall average of errors of the GBM and OUP forecasts.

4. Analysis of Results

Monte Carlo simulations of GBM and OUP models render 1000 different possible paths for TC and RC, respectively, at each time step size considered. Accuracy errors for both annual time steps and averaged monthly time steps, for both GBM and OUP forecasts, were first compared to determine the most accurate time step size for each model. In addition, the LES model outcome for both TC and RC at different timeframes was also calculated and measures of errors for all the three alternative models proposed at an optimum time step were finally compared.

4.1. GBM Forecast. The first 20 of 1000 Monte Carlo paths for the GBM model with a monthly step size using Matlab software may be seen in Figure 2 for both the TC and the RC levels compared to their averaged paths for every 12 values obtained. The tendency for GBM forecasts to steeply increase over time is easily observable in the nonaveraged monthly step paths shown.

The average values of error of all 1000 MC paths obtained through simulation for averaged monthly step and annual-step forecasts are shown in Table 3 for both TC and RC

Table 3: Average of main measures of error for GBM after 1000 MC simulations from 2013 to 2017.

	$\overline{MSE}$	$\overline{MAD}$	$\overline{MAPE}$	$\overline{RMSE}$
TC Averaged Monthly Steps	$5.60x10^3$	46.98	0.50	56.38
TC Annual Steps	$5.20x10^3$	49.02	0.52	58.57
RC Averaged Monthly Steps	58.74	4.55	0.48	5.46
RC Annual Steps	50.85	4.91	0.52	5.84

Table 4: Number of paths for which measures of error are higher for monthly steps than for annual steps in GBM.

	MSE	*MAD*	*MAPE*	*RMSE*
TC	457/1000	455/1000	453/1000	453/1000
RC	439/1000	430/1000	429/1000	429/1000

discounts over the period from 2013 to 2017, which may lead to preliminary conclusions in terms of an accuracy comparison between averaged monthly steps and annual steps.

However, a more exhaustive analysis is shown in Table 4, where the number of times the values of each error measure are higher for monthly steps is expressed over the total number of MC simulations carried out. The results indicate that better values of error are reached the majority of times for averaged monthly step simulations rather than for straight annual ones.

In contrast, short-term accuracy was also evaluated by analysing the error measures of one-year-ahead forecasts (2013) in Table 5 and of two-year-ahead forecasts (2013–2014) in Table 6. The results indicate, as one may expect, that accuracy decreases as the forecasted horizon is widened, with the accuracy of averaged monthly step forecasts remaining higher than annual ones as found above for the five-year forecasting horizon.

4.2. OUP Forecast. Long-term levels for TC and RC, μ, the speed of reversion, λ, and the volatility of the process, σ, are the parameters determined at the model calibration stage (see *Appendix B* for the model calibration explanation.), which define the behaviour of the OUP, shown in Table 7. The calibration was done prior to the Monte Carlo simulation for both TC and RC, with each step size using available historical data from 2004 to 2012. The OUP model was fitted with the corresponding parameters for each case upon MC simulation.

Figure 3 shows the Mean Reversion MC estimations of the TC/RC benchmark values from 2013 to 2017 using monthly steps. The monthly forecasts were rendered from January 2012 through to December 2016 and averaged every twelve values to deliver a benchmark forecast for each year. The averaged results can be comparable to actual data as well as to the annual Monte Carlo simulations following Mean Reversion. The lower-side figures show these yearly averaged monthly step simulation outcomes which clearly move around a dash-dotted red line, indicating the long-term run levels for TC/RC to which they tend to revert.

The accuracy of monthly steps against annual steps for the TC/RC benchmark levels forecast was also tested by determining the number of simulations for which average error measures became higher. Table 8 shows the number of times monthly simulations have been less accurate than annual simulations for five-year-ahead OUP forecasting by comparing the four measures of errors proposed. The results indicate that only 25-32% of the simulations drew a higher average error, which clearly results in a better predictive accuracy for monthly step forecasting of TC/RC annual benchmark levels.

The averaged measures of errors obtained after the MC simulations of the OUP model for both averaged monthly steps and annual steps giving TC/RC benchmark forecasts from 2013 to 2017 are shown in Table 9.

The error levels of the MC simulations shown in Table 9 point towards a higher prediction accuracy of averaged monthly step forecasts of the OUP Model, yielding an averaged MAPE value that is 12.9% lower for TC and RC 5-step-ahead forecasts. In regard to MAPE values, for monthly steps, only 26.6% of the simulations rise above annual MC simulations for TC and 25% for RC 5-step-ahead forecasts, which further underpins the greater accuracy of this OUP set-up for TC/RC level forecasts. A significant lower probability of higher error levels for TC/RC forecasts with monthly MC OUP simulations is reached for the other measures provided. In addition, short-term and long-term prediction accuracy were tested by comparing errors of forecasts for five-year-ahead error measures in Table 10, one-year-ahead in Table 11, and two-year-ahead in Table 12.

With a closer forecasting horizon error, measures show an improvement of forecasting accuracy when average monthly steps are used rather than annual ones. For instance, the MAPE values for 2013 forecast for TC are 68% lower for averaged monthly steps than for annual steps, and also MAPE for 2013–2014 were 30% lower for both TC and RC forecasts. Similarly, better values of error are achieved for the other measures for averaged monthly short-term forecasts than in other scenarios. In addition, as expected, accuracy is increased for closer forecasting horizons where the level of errors shown above becomes lower as the deviation of forecasts is trimmed with short-term predictions.

4.3. LES Forecast. In contrast to GBM and OUP, the LES model lacks any stochastic component, so nondeterministic

Table 5: Average for main measures of error for GBM after 1000 MC simulations for 2013.

	$\overline{MSE}$	$\overline{MAD}$	$\overline{MAPE}$	$\overline{RMSE}$
TC Averaged Monthly Steps	336.58	14.55	0.21	14.55
TC Annual Steps	693.81	20.81	0.30	20.81
RC Averaged Monthly Steps	2.97	1.38	0.20	1.38
RC Annual Steps	6.57	2.05	0.29	2.05

Table 6: Average for main measures of error for GBM after 1000 MC simulations for 2013-2014.

	$\overline{MSE}$	$\overline{MAD}$	$\overline{MAPE}$	$\overline{RMSE}$
TC Averaged Monthly Steps	994.92	21.87	0.31	24.35
TC Annual Steps	$1.33\text{x}10^3$	28.33	0.40	31.04
RC Averaged Monthly Steps	10.35	2.40	0.34	2.71
RC Annual Steps	13.13	2.84	0.34	3.11

Table 7: OUP parameters obtained in the calibration process.

	μ	λ	σ
TC Monthly	63.45	$4.792\text{x}10^5$	2.534
TC Annual	63.45	2.974	1.308
RC Monthly	6.35	$4.763\text{x}10^5$	2.519
RC Annual	6.35	2.972	1.305

methods such as the Monte Carlo are not required to obtain a forecast. Nonetheless, the LES model relies on two *smoothing constants* which must be properly set in order to deliver accurate predictions; hence the values of the smoothing constants were first optimised (see *Appendix B* for LES Model fitting and smoothing constants optimisation.). The optimisation was carried out for one-year-ahead forecasts, two-year-ahead forecasts, and five-year-ahead forecasts by minimising the values of MSE for both TC and RC. The different values used for smoothing constants, as well as the initial values for level and trend obtained by the linear regression of the available dataset from 2004 through to 2012, are shown in Table 12.

Compared one-year-ahead, two-year-ahead, and five-year-ahead LES forecasts for TC and RC are shown in Figure 4, clearly indicating a stronger accuracy for shorter-term forecasts as the *observed* and *forecasted* plotted lines overlap.

Minimum values for error measures achieved through the LES model parameter optimisation are shown in Table 13. The values obtained confirm the strong accuracy for shorter-term forecasts of TC/RC seen in Figure 4.

5. Discussion and Model Comparison

The forecasted values of TC/RC benchmark levels could eventually be applied to broader valuation models for copper concentrates and their trading activities, as well as to the copper smelters' revenue stream, thus the importance of delivering as accurate a prediction as possible in relation to these discounts to make any future application possible. Each of the models presented may be a feasible method on its own with eventual later adaptations to forecasting future values of benchmark TC/RC. Nonetheless, the accuracy of these models as they have been used in this work requires, firstly, a comparison to determine whether any of them could be a good standalone technique and, secondly, to test whether a combination of two or more of them would deliver more precise results.

When comparing the different error measures obtained for all the three models, it is clearly established that results for a randomly chosen simulation of GBM or OUP would be more likely to be more precise had a monthly step been used to deliver annual forecasts instead of an annual-step size. In contrast, average error measures for the entire population of simulations with each step size employed showing that monthly step simulations for GBM and OUP models are always more accurate than straight annual-step forecasts when a shorter time horizon, one or two-year-ahead, is taken into consideration. However, GBM presents a higher level of forecasting accuracy when the average for error measures of all simulations is analysed, employing annual steps for long-term horizons, whereas OUP averaged monthly step forecasts remain more accurate when predicting long-term horizons. Table 14 shows the error improvement for the averaged monthly step forecasts of each model. Negative values indicate that better levels of error averages have been found in straight annual forecasts than for monthly step simulations.

Considering the best results for each model and comparing their corresponding error measures, we can opt for the best technique to employ among the three proposed in this paper. Hence, GBM delivers the most accurate one-year forecast when averaging the next twelve-month predictions for TC/RC values, as does the MR-OUP model. Table 15 shows best error measures for one-year-ahead forecasts for GBM, OUP, and LES models.

Unarguably, the LES model generates minimal error measures for one-year-ahead forecasts, significantly less than the other models employed. A similar situation is found for two-year-ahead forecasts where minimum error measures are also delivered by the LES model, shown in Table 16.

Finally, accuracy measures for five-year-ahead forecasts of the GBM model might result in somewhat contradictory

Table 8: Number of paths for which measures of error are higher for monthly steps than for annual steps in MR-OUP 5-Steps MC simulation.

	MSE	*MAD*	*MAPE*	*RMSE*
TC	311/1000	283/1000	266/1000	266/1000
RC	316/1000	281/1000	250/1000	250/1000

Table 9: Average for main measures of error for MR-OUP after 1000 MC simulations 2013-2017.

	$\overline{MSE}$	$\overline{MAD}$	$\overline{MAPE}$	$\overline{RMSE}$
TC Averaged Monthly Steps	842.54	26.14	0.27	28.95
TC Annual Steps	$1.14\text{x}10^{3}$	29.31	0.31	33.25
RC Averaged Monthly Steps	8.40	2.61	0.27	2.89
RC Annual Steps	11.48	2.94	0.31	3.33

Table 10: Average for main measures of error for MR-OUP after 1000 MC simulations 2013.

	$\overline{MSE}$	$\overline{MAD}$	$\overline{MAPE}$	$\overline{RMSE}$
TC Averaged Monthly Steps	39.14	5.22	0.07	5.22
TC Annual Steps	366.43	15.55	0.22	15.55
RC Averaged Monthly Steps	0.39	0.51	0.07	0.51
RC Annual Steps	3.63	1.54	0.22	1.54

Table 11: Average for main measures of error for MR-OUP after 1000 MC simulations 2013-2014.

	$\overline{MSE}$	$\overline{MAD}$	$\overline{MAPE}$	$\overline{RMSE}$
TC Averaged Monthly Steps	371.65	15.67	0.18	19.00
TC Annual Steps	682.44	21.85	0.26	23.24
RC Averaged Monthly Steps	3.76	1.57	0.18	1.91
RC Annual Steps	6.89	2.19	0.26	2.43

Table 12: LES Model Optimised Parameters.

	L_0	T_0	α	β
TC 1 year	71.3611	-1.5833	-0.2372	0.1598
RC 1 year	7.1333	-0.1567	-0.2368	0.1591
TC 2 years	71.3611	-1.5833	$-1.477\text{x}10^{-4}$	777.4226
RC 2 years	7.1333	-0.1567	$-1.448\text{x}10^{-4}$	789.8336
TC 5 years	71.3611	-1.5833	-0.2813	0.1880
RC 5 years	7.1333	-0.1567	-0.2808	0.1880

Table 13: LES error measures for different steps-ahead forecasts.

	$\overline{MSE}$	$\overline{MAD}$	$\overline{MAPE}$	$\overline{RMSE}$
TC 1 year	$1.0746\text{x}10^{-5}$	0.0033	$4.6831\text{x}10^{-5}$	0.0033
RC 1 year	$5.1462\text{x}10^{-8}$	$2.2685\text{x}10^{-4}$	$3.2408\text{x}10^{-5}$	$2.2685\text{x}10^{-4}$
TC 2 years	189.247	9.7275	0.1057	13.7567
RC 2 years	1.8977	0.9741	0.1059	1.3776
TC 5 years	177.5531	7.8881	0.0951	10.8422
RC 5 years	1.1759	0.7889	0.0952	1.0844

terms for MSE reaching better values for annual steps than for averaged monthly steps, while the other figures do better on averaged monthly steps. Addressing the definition of MSE, this includes the variance of the estimator as well as its bias, being equal to its variance in the case of unbiased estimators. Therefore, MSE measures the quality of the estimator but also magnifies estimator deviations from actual values since both positive and negative values are squared and averaged. In contrast, RMSE is calculated as the square root of MSE and, following the previous analogy, stands for the standard

Table 14: Error Average Improvement for averaged monthly steps before annual steps.

	$\overline{MSE}$	$\overline{MAD}$	$\overline{MAPE}$	$\overline{RMSE}$
GBM TC/RC 1 Year	53.26%/56.53%	32.72%/35.87%	33.33%/35.48%	32.72%/35.87%
GBM TC/RC 2 Years	26.95%/26.35%	25.75%/15.97%	26.19%/2.86%	24.34%/13.69%
GBM TC/RC 5 Years	-11.73%/-11.04%	8.73%/7.09%	9.26%/7.69%	7.47%/5.65%
OUP TC/RC 1 Year	88.08%/87.08%	63.71%/62.50%	62.91%/63.19%	63.68%/62.24%
OUP TC/RC 2 Years	46.28%/44.21%	27.71%/26.17%	29.99%/30.75%	22.16%/20.75%
OUP TC/RC 5 Years	26.02%/25.53%	10.93%/9.97%	13.12%/12.18%	13.06%/12.69%

Table 15: TC/RC best error measures for 2013 forecasts after 1000 MC simulations.

	$\overline{MSE}$	$\overline{MAD}$	$\overline{MAPE}$	$\overline{RMSE}$
GBM TC/RC 1 Year∗	331.11/3.36	14.25/1.43	0.20/0.20	14.25/1.43
OUP TC/RC 1 Year∗	41.05/0.42	5.36/0.54	0.0766/0.0777	5.36/0.54
LES TC/RC 1 Year	$1.07\text{x}10^{-5}/5.14\text{x}10^{-8}$	$0.003/2.26\text{x}10^{-4}$	$4.68\text{x}10^{-5}/3.24\text{x}10^{-5}$	$0.003/2.26\text{x}10^{-4}$

∗Averaged monthly steps.

Table 16: TC/RC Best error measures for 2013–2014 forecasts after 1000 MC simulations.

	$\overline{MSE}$	$\overline{MAD}$	$\overline{MAPE}$	$\overline{RMSE}$
GBM TC/RC 2 Steps∗	$1.03\text{x}10^{3}/10.06$	22.00/2.42	0.31/0.34	24.47/2.71
OUP TC/RC 2 Steps∗	373.31/3.76	15.73/1.58	0.1802/0.1813	19.00/1.91
LES TC/RC 2 Steps	189.24/1.89	9.72/0.97	0.1057/0.1059	13.75/1.37

∗Averaged monthly steps.

deviation of the estimator if MSE were considered to be the variance. Though RMSE overreacts when high values of MSE are reached, it is less prone to this than MSE since it is calculated as its squared root, thus not accounting for large errors as disproportionately as MSE does. Furthermore, as we have compared an average of 1000 measures of errors corresponding to each MC simulation performed, the values obtained for average RMSE stay below the square root of average MSE, which indicates that some of these disproportionate error measures are, to some extent, distorting the latter. Hence, RMSE average values point towards a higher accuracy for GBM five-year forecasts with averaged monthly steps, which is further endorsed by the average values of MAD and MAPE, thus being the one used for comparison with the other two models as shown in Table 17.

The final comparison clearly shows how the LES model outperforms the other two at all average measures provided, followed by the OUP model in accuracy, although the latter more than doubles the average MAPE value for LES.

The results of simulations indicate that measures of errors tend to either differ slightly or not at all for either forecasts of any timeframe. A coherent value with the 10:1 relation can then be given with close to the same level of accuracy by multiplying RC forecasts or dividing TC ones by 10.

6. Conclusions

Copper TC/RC are a keystone for pricing copper concentrates which are the actual feedstock for copper smelters. The potential evolution of TC/RC is a question of both economic and technical significance for miners, as their value decreases the potential final selling price of concentrates. Additionally, copper miners' revenues are more narrowly related to the market price of copper, as well as to other technical factors such as ore dilution or the grade of the concentrates produced. Smelters, on the contrary, are hugely affected by the discount which they succeed in getting when purchasing the concentrates, since that makes up the largest part of their gross revenue, besides other secondary sources. In addition, eventual differences between TC/RC may give commodity traders ludicrous arbitrage opportunities. Also, differences between short- and long-term TC/RC agreements offer arbitrage opportunities for traders, hence comprising a part of their revenue in the copper concentrate trading business, as well copper price fluctuations and the capacity to make economically optimum copper blends.

As far as we are aware, no rigorous research has been carried out on the behaviour of these discounts. Based on historical data on TC/RC agreed upon in the LME Copper Week from 2004 to 2017, three potentially suitable forecasting models for TC/RC annual benchmark values have been compared through four measures of forecasting accuracy at different horizons. These models were chosen, firstly, due to their broad implementation and proven capacity in commodity prices forecasting that they all share and, secondly, because of the core differences in terms of price behaviour with each other.

Focusing on the MAPE values achieved, those obtained by the LES model when TC and RC are treated independently have been significantly lower than for the rest of the models. Indeed, one-year-ahead MAPE measures for TC values for the GBM model (20%) almost triple those of

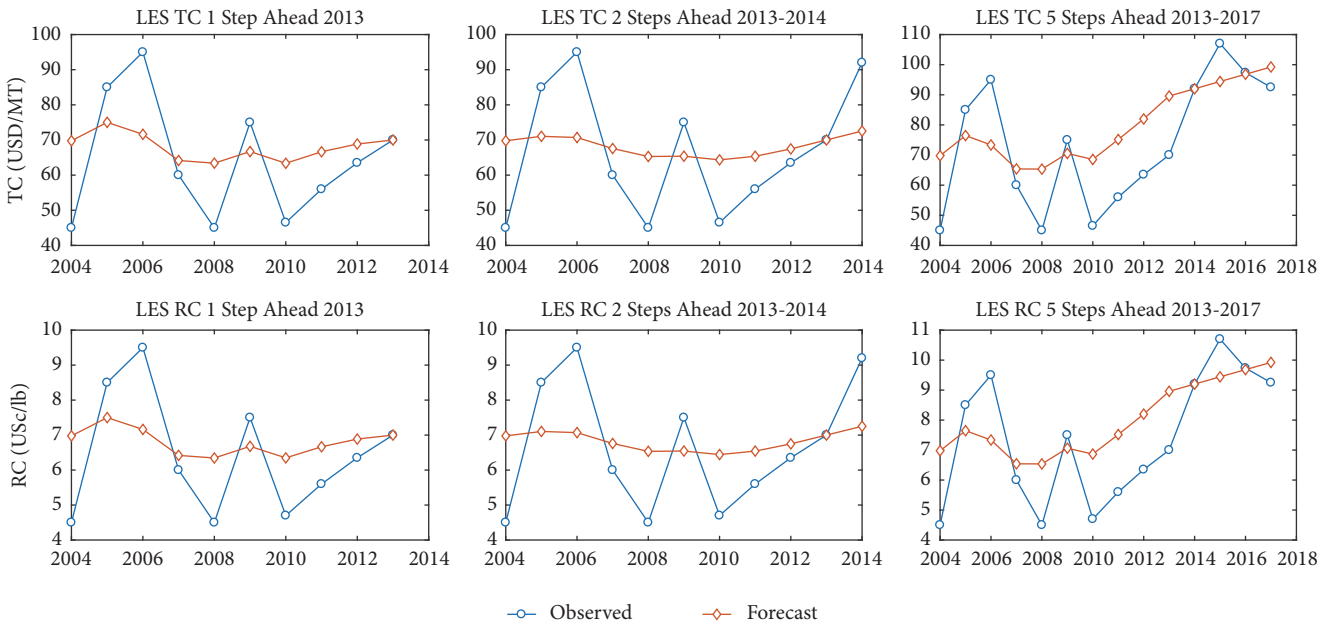


Figure 4: One-step-ahead forecasts (2013), two-step-ahead forecasts (2013-2014) and five-step-ahead forecasts (2013-2017) for TC and RC using the Linear Exponential Smoothing model (LES).

Table 17: TC/RC Best error measures for 2013–2017 forecasts after 1000 MC simulations.

	$\overline{MSE}$	$\overline{MAD}$	$\overline{MAPE}$	$\overline{RMSE}$
GBM TC/RC 5 Steps∗	$6.00x10^3$/61.65	46.51/4.59	0.49/0.48	55.71/5.51
OUP TC/RC 5 Steps∗	843.34/8.43	26.16/2.62	0.2715/0.2719	28.96/2.89
LES TC/RC 5 Steps	177.55/1.17	7.88/0.78	0.0951/0.0952	10.84/1.08

∗Averaged monthly steps.

the OUP model (7.66%), in contrast with the significantly lower values from the LES model (0.0046%). This gap tends to be narrowed when TC/RC values are forecasted at longer horizons, when most measures of error become more even. The GBM and OUP models have proven to deliver better accuracy performance when the TC/RC values are projected monthly and then averaged to obtain annual benchmark forecasts. Even so, the LES model remains the most accurate of all with MAPE values of 10% at two-year-ahead forecasts, with 18% and 31% for TC for OUP and GBM, respectively.

Finally, despite TC and RC being two independent discounts applied to copper concentrates, they are both set jointly with an often 10:1 relation as our data reveals. This relation also transcends to simulation results and error measures, hence showing a negligible discrepancy between the independent forecasting of TC and RC, or the joint forecasting of both values, keeping the 10:1 relation. This is, for instance, the case of the five-year-ahead OUP MAPE values (0.2715/0.2719) which were obtained without observing the 10:1 relation in the data. A similar level of discrepancy was obtained at any horizon with any model, which indicates that both values could be forecasted with the same accuracy using the selected model with any of them and then applying the 10:1 relation.

Our findings thus suggest that both at short and at long-term horizons TC/RC annual benchmark levels tend to exhibit a pattern which is best fit by an LES model. This indicates that these critical discounts for the copper trading business do maintain a certain dependency on past values. This would also suggest the existence of cyclical patterns in copper TC/RC, possibly driven by many of the same market factors that move the price of copper.

This work contributes by delivering a formal tool for smelters or miners to make accurate forecasts of TC/RC benchmark levels. The level of errors attained indicates the LES model may be a valid model to forecast these crucial discounts for the copper market. In addition, our work further contributes to helping market participants to project the price of concentrates with an acceptable degree of uncertainty, as now they may include a fundamental element for their estimation. This would enable them to optimise the way they produce or process these copper concentrates. Also, the precise knowledge of these discounts' expected behaviour contributes to letting miners, traders, and smelters alike take the maximum advantage from the copper concentrate trading agreements that they are part of. As an additional contribution, this work may well be applied to gold or silver RC, which are relevant deduction concentrates when these have a significant amount of gold or silver.

Once TC/RC annual benchmark levels are able to be forecasted with a certain level of accuracy, future research should go further into this research through the exploration of the potential impact that other market factors may have on these discounts.

As a limitation of this research, we should point out the timespan of the data considered, compared to those of other forecasting works, on commodity prices for example, which use broader timespans. For our case, we have considered the maximum available sufficiently reliable data on TC/RC benchmark levels, starting back in 2004, as there is no reliable data beyond this year. Also, as another limitation, we have used four measures of error which are among the most frequently used to compare the accuracy of different models. However, other measures could have been used at an individual level to test each model's accuracy.

Appendix

A. Models

A.1. Geometric Brownian Motion (GBM). GBM can be written as a generalisation of a Wiener (continuous time-stochastic Markov process, with independent increments and whose changes over any infinite interval of time are normally distributed, with a variance that increases linearly with the time interval [22].) process:

$$dx = \alpha x dt + \sigma x dz \tag{A.1}$$

where according to Marathe and Ryan [30] the first term is known as the Expected Value, whereas the second is the Stochastic component, with α being the drift parameter and σ the volatility of the process. Also, dz is the Wiener process which induces the abovementioned stochastic behaviour in the model:

$$dz = \epsilon_t \sqrt{dt} \longrightarrow \epsilon_t \sim N(0, 1) \tag{A.2}$$

The GBM model can be expressed in discrete terms according to

$$\Delta x = x_t - x_{t-1} = \alpha x_{t-1} \Delta t + \sigma x_{t-1} \epsilon \sqrt{\Delta t} \tag{A.3}$$

In GBM percentage changes in x, $\Delta x/x$ are normally distributed; thus absolute changes in x, Δx are *lognormally* distributed. Also, the expected value and variance for $x(t)$ are

$$\mathbb{E}[x(t)] = x_0 e^{\alpha t} \tag{A.4}$$

$$\text{var}[x(t)] = x_0^2 e^{2\alpha t} \left(e^{\sigma^2 t} - 1\right) \tag{A.5}$$

A.2. Orstein-Uhlenbeck Process (OUP). The OUP process was first defined by Uhlenbeck and Orstein [31] as an alternative to the regular Brownian Motion to model the velocity of the diffusion movement of a particle that accounts for its losses due to friction with other particles. The OUP process can be regarded as a modification of Brownian Motion in continuous time where its properties have been changed (Stationary, Gaussian, Markov, and stochastic process.) [32]. These modifications cause the process to move towards a central position, with a stronger attraction the further it is from this position. As mentioned above, the OUP is usually employed to model commodity prices and is the simplest version of a *mean-reverting process* [22]:

$$dS = \lambda(\mu - S) dt + \sigma dW_t \tag{A.6}$$

where S is the level of prices, μ the long-term average to which prices tend to revert, and λ the speed of reversion. Additionally, in a similar fashion to that of the GBM, σ is the volatility of the process and dW_t is a Wiener process with an identical definition. However, in contrast to GBM, time intervals in OUP are not independent since differences between current levels of prices, S, and long-term average prices, μ, make the expected change in prices, dS, more likely either positive or negative.

The discrete version of the model can be expressed as follows:

$$S_t = \mu\left(1 - e^{-\lambda \Delta t}\right) + e^{-\lambda \Delta t} S_{t-1} + \sigma \sqrt{\frac{1 - e^{-2\lambda \Delta t}}{2\lambda}} dW_t \tag{A.7}$$

where the expected value for $x(t)$ and the variance for $(x(t) - \mu)$ are

$$\mathbb{E}[x(t)] = \mu + (x_0 - \mu) e^{-\lambda t} \tag{A.8}$$

$$\text{var}[x(t) - \mu] = \frac{\sigma^2}{2\lambda}\left(1 - e^{-2\lambda t}\right) \tag{A.9}$$

It can be derived from previous equations that as time increases prices will tend to long-term average levels, μ. In addition, with large time spans, if the speed of reversion, λ, becomes high, variance tends to 0. On the other hand, if the speed of reversion is 0 then $\text{var}[x(t)] \longrightarrow \sigma^2 t$, making the process a simple Brownian Motion.

A.3. Holt's Linear Exponential Smoothing (LES). Linear Exponential Smoothing models are capable of considering both levels and trends at every instant, assigning higher weights in the overall calculation to values closer to the present than to older ones. LES models carry that out by constantly updating local estimations of levels and trends with the intervention of one or two smoothing constants which enable the models to dampen older value effects. Although it is possible to employ a single smoothing constant for both the level and the trend, known as Brown's LES, to use two, one for each, known as Holt's LES, is usually preferred since Brown's LES tends to render estimations of the trend "unstable" as suggested by authors such as Nau [33]. Holt's LES model comes defined by the level, trend, and forecast updating equations, each of these expressed as follows, respectively:

$$L_t = \alpha Y_t + (1 - \alpha)\left(L_{t-1} + T_{t-1}\right) \tag{A.10}$$

$$T_t = \beta\left(L_t - L_{t-1}\right) + (1 - \beta) T_{t-1} \tag{A.11}$$

$$\widehat{Y}_{t+k} = L_t + kT_t \tag{A.12}$$

With α being the first smoothing constant for the levels and β the second smoothing constant for the trend. Higher values for the smoothing constants imply that either levels or trends are changing rapidly over time, whereas lower values imply the contrary. Hence, the higher the constant, the more uncertain the future is believed to be.

B. Model Calibration

Each model has been calibrated individually using two sets of data containing TC and RC levels, respectively, from 2004 to 2012. The available dataset is comprised of data from 2004 to 2017, which make the calibration data approximately 2/3 of the total.

B.1. GBM. Increments in the logarithm of variable x are distributed as follows:

$$\Delta(\ln x) \sim N\left(\left(\alpha - \frac{\sigma^2}{2}\right)t, \sigma t\right) \tag{B.1}$$

Hence, if m is defined as the sample mean of the difference of the natural logarithm of the time series for TC/RC levels considered for the calibration and n as the number of increments of the series considered, with n=9,

$$m = \frac{1}{n}\sum_{t=1}^{n}\left(\ln x_t - \ln x_{t-1}\right) \tag{B.2}$$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\ln x_t - \ln x_{t-1} - m\right)^2} \tag{B.3}$$

$$m = \alpha - \frac{\sigma^2}{2} \tag{B.4}$$

$$s = \sigma \tag{B.5}$$

B.2. OUP. The OUP process is an AR1 process [22] whose resolution is well presented by Woolridge [34] using *OLS* techniques, fitting the following:

$$y_t = a + by_{t-1} + \varepsilon_t \tag{B.6}$$

Hence, the estimators for the parameters of the OUP model are obtained by OLS for both TC and RC levels independently:

$$\widehat{\lambda} = -\frac{\ln b}{\Delta t} \tag{B.7}$$

$$\widehat{\mu} = \frac{a}{1-b} \tag{B.8}$$

$$\widehat{\sigma} = \varepsilon_t\sqrt{\frac{2\ln(1+b)}{(1+b)^2 - 1}} \tag{B.9}$$

B.3. LES. A linear regression is conducted on the input dataset available to find the starting parameters for the LES model, the initial Level, L_0, and the initial value of the trend, T_0, irrespective of TC values and RC values. Here, as recommended by Gardner [35], the use of OLS is highly advisable due to the erratic behaviour shown by trends in the historic data, so the obtaining of negative values of S_0 is prevented. Linear regression fulfils the following:

$$Y_t = at + b \tag{B.10}$$

$$L_0 = b \tag{B.11}$$

$$T_0 = a \tag{B.12}$$

By fixing the two smoothing constants, the values for the forecasts, $\widehat{Y}_{t+k}$, can be calculated at each step using the model equations. There are multiple references in the literature on what the optimum range for each smoothing constant is; Gardner [35] speaks of setting moderate values for both parameter less than 0.3 to obtain the best results. Examples pointing out the same may be found in Brown [36], Coutie [37], Harrison [38], and Montgomery and Johnson [39]. Also, for many applications, Makridakis and Hibon [20] and Chatfield [40] found that parameter values should fall within the range of 0.3-1. On the other hand, McClain and Thomas [41] provided a condition of stability for the nonseasonal *LES* model given by

$$\begin{aligned} 0 < \alpha < 2 \\ 0 < \beta < \frac{4 - 2\alpha}{\alpha} \end{aligned} \tag{B.13}$$

Also, the largest possible value of α that allows the avoidance of areas of oscillation is proposed by McClain and Thomas [41] and McClain [42]:

$$\alpha < \frac{4\beta}{(1+\beta)^2} \tag{B.14}$$

However, according to Gardner, there is no tangible proof that this value improves accuracy in any form. Nonetheless, we have opted to follow what Nau [33] refers to as *"the usual way"*, namely, minimising the *Mean Squared Error (MSE)* of the one-step-ahead forecast of TC/RC for each input data series previously mentioned. To do so, Matlab's *fminsearch* function has been used with function and variable tolerance levels of $1x10^{-4}$ as well as a set maximum number of function iterations and function evaluations of $1x10^{6}$ to limit computing resources. In Table 18 the actual number of necessary iterations to obtain optimum values for smoothing constants is shown. As can be seen, the criteria are well beyond the final results, which ensured that an optimum solution was reached with assumable computing usage (the simulation required less than one minute) and with a high degree of certainty.

Table 18: Iterations on LES smoothing constants optimisation.

	Iterations	Function Evaluations
TC 1-Step	36	40
TC 2-Steps	249945	454211
TC 5-Steps	40	77
RC 1-Step	32	62
RC 2-Steps	254388	462226
RC 5-Steps	34	66

C. Error Calculations

C.1. Mean Square Error (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{Y}_i - Y_i\right)^2 \quad \text{(C.1)}$$

where $\widehat{Y}_i$ are the forecasted values and Y_i those observed.

C.2. Mean Absolute Deviation (MAD)

$$MAD = \frac{1}{n}\sum_{i=1}^{n}\left|\widehat{Y}_i - \overline{Y}\right| \quad \text{(C.2)}$$

where $\widehat{Y}_i$ are the forecasted values and $\overline{Y}$ the average value of all the observations.

C.3. Mean Absolute Percentage Error (MAPE)

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|\widehat{Y}_i - Y_i\right|}{Y_i} \quad \text{(C.3)}$$

The above formula is expressed in parts-per-one and is the one used in the calculations conducted here. Hence, multiplying the result by 100 would deliver percentage outcomes.

C.4. Root Mean Square Error (RMSE)

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{Y}_i - Y_i\right)^2} \quad \text{(C.4)}$$

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Supplementary Materials

The Matlab codes developed to carry out the simulations for each of the models proposed, as well as the datasets used and the Matlab workspaces with the simulations outcomes as they have been reflected in this paper, are provided in separate folders for each model. GBM folder contains the code to make the simulations for the Geometric Brownian Motion model, "GBM.m", as well as a separate script which allows determining the total number of monthly step simulations for which errors levels have been lower than annual-step simulations, "ErrCouter.m". The datasets used are included in two excel files: "TC.xlsx" and "RC.xlsx". "GBM_OK_MonthlySteps.mat" is the Matlab workspace containing the outcome of monthly step simulations, whereas "GBM_OK_AnnualSteps.mat" contains the outcome for annual-step simulations. Also, "GBM_OK.mat" is the workspace to be loaded prior to performing GBM simulations. MR folder contains the code file to carry out the simulations for the Orstein-Uhlenbeck Process, "MR.m", as well as the separate script to compare errors of monthly step simulations with annual-step simulations, "ErrCounter.m". "TC.xlsx" and "RC.xlsx" contain the TC/RC benchmark levels from 2004 to 2017. Finally, monthly step simulations' outcome has been saved in the workspace "MR_OK_MonthlySteps.mat", while annual-step simulations' outcome has been saved in the Matlab workspace "MR_OK_AnnualSteps.mat". Also, "MR_OK.mat" is the workspace to be loaded prior to performing MR simulations. LES folder contains the code file to carry out the calculations necessary to obtain the LES model forecasts, "LES.m". Three separate Matlab functions are included: "mapeLES.m", "mapeLES1.m", and "mapeLES2.m", which define the calculation of the MAPE for the LES' five-year-ahead, one-year-ahead, and two-year-ahead forecasts. Also, "LES_OK.mat" is the workspace to be loaded prior to performing LES simulations. Finally, graphs of each model have been included in separate files as well, in a subfolder named "Graphs" within each of the models' folder. Figures are included in Matlab's format (.fig) and TIFF format. TIFF figures have been generated in both colours and black and white for easier visualization. *(Supplementary Materials)*

References

[1] United Nations, *International Trade Statistics Yearbook*, vol. 2, 2016.

[2] International Copper Study Group, *The World Copper Factbook*, 2017.

[3] K. M. Johnson, J. M. Hammarstrom, M. L. Zientek, and C. L. Dicken, "Estimate of undiscovered copper resources of the world, 2013," U.S. Geological Survey, 2017.

[4] M. E. Schlesinger and W. G. Davenport, *Extractive Metallurgy of Copper*, Elsevier, 5th edition, 2011.

[5] S. Glöser, M. Soulier, and L. A. T. Espinoza, "Dynamic analysis of global copper flows. Global stocks, postconsumer material flows, recycling indicators, and uncertainty evaluation," *Environmental Science & Technology*, vol. 47, no. 12, pp. 6564–6572, 2013.

[6] The London Metal Exchange, "Special contract rules for copper grade A," https://www.lme.com/en-GB/Trading/Brands/Composition.

[7] U. Soderstrom, "Copper smelter revenue stream," 2008.

[8] F. K. Crundwell, *Finance for Engineers (Evaluation and Funding of Capital Projects)*, 2008.

[9] Metal Bulletin TC/RC Index, "Meteoritical Bulletin," https://www.metalbulletin.com/Article/3319574/Copper/PRICING-NOTICE-Specification-of-Metal-Bulletin-copper-concentrates-TCRC-index.html.

[10] J. Savolainen, "Real options in metal mining project valuation: review of literature," *Resources Policy*, vol. 50, pp. 49–65, 2016.

[11] W. J. Hahn, J. A. DiLellio, and J. S. Dyer, "Risk premia in commodity price forecasts and their impact on valuation," *Energy Econ*, vol. 7, p. 28, 2018.

[12] C. Alberto, C. Pinheiro, and V. Senna, "Price forecasting through multivariate spectral analysis: evidence for commodities of BM&Fbovespa," *Brazilian Business Review*, vol. 13, no. 5, pp. 129–157, 2016.

[13] Z. Hloušková, P. Ženíšková, and M. Prášilová, "Comparison of agricultural costs prediction approaches," *AGRIS on-line Papers in Economics and Informatics*, vol. 10, no. 1, pp. 3–13, 2018.

[14] T. Xiong, C. Li, Y. Bao, Z. Hu, and L. Zhang, "A combination method for interval forecasting of agricultural commodity futures prices," *Knowledge-Based Systems*, vol. 77, pp. 92–102, 2015.

[15] Y. E. Shao and J.-T. Dai, "Integrated feature selection of ARIMA with computational intelligence approaches for food crop price prediction," *Complexity*, vol. 2018, Article ID 1910520, 17 pages, 2018.

[16] R. Heaney, "Does knowledge of the cost of carry model improve commodity futures price forecasting ability? A case study using the london metal exchange lead contract," *International Journal of Forecasting*, vol. 18, no. 1, pp. 45–65, 2002.

[17] S. Shafiee and E. Topal, "An overview of global gold market and gold price forecasting," *Resources Policy*, vol. 35, no. 3, pp. 178–189, 2010.

[18] Y. Li, N. Hu, G. Li, and X. Yao, "Forecasting mineral commodity prices with ARIMA-Markov chain," in *Proceedings of the 2012 4th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, pp. 49–52, Nanchang, China, August 2012.

[19] J. V. Issler, C. Rodrigues, and R. Burjack, "Using common features to understand the behavior of metal-commodity prices and forecast them at different horizons," *Journal of International Money and Finance*, vol. 42, pp. 310–335, 2014.

[20] S. Makridakis, M. Hibon, and C. Moser, "Accuracy of forecasting: an empirical investigation," *Journal of the Royal Statistical Society. Series A (General)*, vol. 142, no. 2, p. 97, 1979.

[21] S. Makridakis, A. Andersen, R. Carbone et al., "The accuracy of extrapolation (time series) methods: Results of a forecasting competition," *Journal of Forecasting*, vol. 1, no. 2, pp. 111–153, 1982.

[22] A. K. Dixit and R. S. Pindyck, *Investment under uncertainty*, vol. 256, 1994.

[23] F. Black and M. Scholes, "The pricing of options corporate liabilities," *Journal of Political Economy*, vol. 81, pp. 637–659, 1973.

[24] N. Foo, H. Bloch, and R. Salim, "The optimisation rule for investment in mining projects," *Resources Policy*, vol. 55, pp. 123–132, 2018.

[25] P. Kalekar, "Time series forecasting using Holt-Winters exponential smoothing," *Kanwal Rekhi School of Information Technology*, Article ID 04329008, pp. 1–13, 2004, http://www.it.iitb.ac.in/~praj/acads/seminar/04329008_Exponential-Smoothing.pdf.

[26] L. Gómez-Valle and J. Martínez-Rodríguez, "Advances in pricing commodity futures: multifactor models," *Mathematical and Computer Modelling*, vol. 57, no. 7-8, pp. 1722–1731, 2013.

[27] M. Khashei and M. Bijari, "A novel hybridization of artificial neural networks and ARIMA models for time series forecasting," *Applied Soft Computing*, vol. 11, no. 2, pp. 2664–2675, 2011.

[28] M. Khashei and M. Bijari, "An artificial neural network (p, d, q) model for timeseries forecasting," *Expert Systems with Applications*, vol. 37, no. 1, pp. 479–489, 2010.

[29] A. Choudhury and J. Jones, "Crop yield prediction using time series models," *The Journal of Economic Education*, vol. 15, pp. 53–68, 2014.

[30] R. R. Marathe and S. M. Ryan, "On the validity of the geometric brownian motion assumption," *The Engineering Economist*, vol. 50, no. 2, pp. 159–192, 2005.

[31] G. E. Uhlenbeck and L. S. Ornstein, "On the theory of the Brownian motion," *Physical Review A: Atomic, Molecular and Optical Physics*, vol. 36, no. 5, pp. 823–841, 1930.

[32] Á. Tresierra Tanaka and C. M. Carrasco Montero, "Valorización de opciones reales: modelo ornstein-uhlenbeck," *Journal of Economics, Finance and Administrative Science*, vol. 21, no. 41, pp. 56–62, 2016.

[33] R. Nau, *Forecasting with Moving Averages*, Duke's Fuqua School of Business, Duke University, Durham, NC, USA, 2014, http://people.duke.edu/128.~rnau/Notes_on_forecasting_with_moving_averages–Robert_Nau.pdf.

[34] J. M. Woolridge, *Introductory Econometrics*, 2011.

[35] E. S. Gardner Jr., "Exponential smoothing: the state of the art-part II," *International Journal of Forecasting*, vol. 22, no. 4, pp. 637–666, 2006.

[36] R. G. Brown, *Smoothing, Forecasting and Prediction of Discrete Time Series*, Dover Phoenix Editions, Dover Publications, 2nd edition, 1963.

[37] G. Coutie et al., *Short-Term Forecasting, ICT Monogr. No. 2*, Oliver Boyd, Edinburgh, Scotland, 1964.

[38] P. J. Harrison, "Exponential smoothing and short-term sales forecasting," *Management Science*, vol. 13, no. 11, pp. 821–842, 1967.

[39] D. C. Montgomery and L. A. Johnson, *Forecasting and Time Series Analysis*, New York, NY, USA, 1976.

[40] C. Chatfield, "The holt-winters forecasting procedure," *Journal of Applied Statistics*, vol. 27, no. 3, p. 264, 1978.

[41] J. O. McClain and L. J. Thomas, "Response-variance tradeoffs in adaptive forecasting," *Operations Research*, vol. 21, pp. 554–568, 1973.

[42] J. O. McClain, "Dynamics of exponential smoothing with trend and seasonal terms," *Management Science*, vol. 20, no. 9, pp. 1300–1304, 1974.

[43] J. V. Issler and L. R. Lima, "A panel data approach to economic forecasting: the bias-corrected average forecast," *Journal of Econometrics*, vol. 152, no. 2, pp. 153–164, 2009.

[44] M. F. Hamid and A. Shabri, "Palm oil price forecasting model: an autoregressive distributed lag (ARDL) approach," in *Proceedings of the 3rd ISM International Statistical Conference 2016 (ISM-III): Bringing Professionalism and Prestige in Statistics*, Kuala Lumpur, Malaysia, 2017.

[45] H. Duan, G. R. Lei, and K. Shao, "Forecasting crude oil consumption in China using a grey prediction model with an optimal fractional-order accumulating operator," *Complexity*, vol. 2018, 12 pages, 2018.

[46] M. J. Brennan and E. S. Schwartz, "Evaluating natural resource investments," *The Journal of Business*, vol. 58, no. 2, pp. 135–157, 1985.

[47] R. McDonald and D. Siegel, "The value of waiting to invest," *The Quarterly Journal of Economics*, vol. 101, no. 4, pp. 707–727, 1986.

[48] K. Zhang, A. Nieto, and A. N. Kleit, "The real option value of mining operations using mean-reverting commodity prices," *Mineral Economics*, vol. 28, no. 1-2, pp. 11–22, 2015.

[49] R. K. Sharma, "Forecasting gold price with box jenkins autoregressive integrated moving average method," *Journal of International Economics*, vol. 7, pp. 32–60, 2016.

[50] J. Johansson, "Boliden capital markets day 2007," 2007.

[51] N. Svante, "Boliden capital markets day September 2011," 2011.

[52] Teck, "Teck modelling workshop 2012," 2012.

[53] A. Shaw, "AngloAmerican copper market outlook," in *Proceedings of the Copper Conference Brisbane*, vol. 24, Nanning, China, 2015, http://www.minexconsulting.com/publications/CRU.

[54] P. Willbrandt and E. Faust, "Aurubis DVFA analyst conference report 2013," 2013.

[55] P. Willbrandt and E. Faust, "Aurubis DVFA analyst conference report 2014," 2014.

[56] A. G. Aurubis, "Aurubis copper mail December 2015," 2015.

[57] D. B. Drouven and E. Faust, "Aurubis DVFA analyst conference report 2015," 2015.

[58] A. G. Aurubis, "Aurubis copper mail February 2016," 2016.

[59] EY, "EY mining and metals commodity briefcase," 2017.

[60] J. Schachler, *Aurubis AG interim report 2017*, Frankfurt, Germany, 2017.

[61] Y. Nakazato, "Aurubis AG interim report," 2017.

Hindawi
Complexity
Volume 2019, Article ID 6352657, 12 pages
https://doi.org/10.1155/2019/6352657

Research Article

The Bass Diffusion Model on Finite Barabasi-Albert Networks

M. L. Bertotti and G. Modanese

Free University of Bolzano-Bozen, Faculty of Science and Technology, 39100 Bolzano, Italy

Correspondence should be addressed to G. Modanese; Giovanni.Modanese@unibz.it

Received 14 November 2018; Accepted 3 March 2019; Published 1 April 2019

Guest Editor: Marisol B. Correia

Using a heterogeneous mean-field network formulation of the Bass innovation diffusion model and recent exact results on the degree correlations of Barabasi-Albert networks, we compute the times of the diffusion peak and compare them with those on scale-free networks which have the same scale-free exponent but different assortativity properties. We compare our results with those obtained for the SIS epidemic model with the spectral method applied to adjacency matrices. It turns out that diffusion times on finite Barabasi-Albert networks are at a minimum. This may be due to a little-known property of these networks: whereas the value of the assortativity coefficient is close to zero, they look disassortative if one considers only a bounded range of degrees, including the smallest ones, and slightly assortative on the range of the higher degrees. We also find that if the trickle-down character of the diffusion process is enhanced by a larger initial stimulus on the hubs (via a inhomogeneous linear term in the Bass model), the relative difference between the diffusion times for BA networks and uncorrelated networks is even larger, reaching, for instance, the 34% in a typical case on a network with 10^4 nodes.

1. Introduction

The study of epidemic diffusion is one of the most important applications of network theory [1–3], the absence of an epidemic threshold on scale-free networks being perhaps the best known result [4]. This result essentially also holds for networks with degree correlations [5], although some exceptions have been pointed out [6, 7]. In [8] the dependence of the epidemic threshold and diffusion time on the network assortativity was investigated, using a degree-preserving rewiring procedure which starts from a Barabasi-Albert network and analysing the spectral properties of the resulting adjacency matrices. In this paper we mainly focus on Barabasi-Albert (BA) networks [9], using the exact results by Fotouhi and Rabbat on the degree correlations [10]. We employ the mean-field method [11] and our network formulation of the Bass diffusion model for the description of the innovation diffusion process [12, 13]. We solve numerically the equations and find the time of the diffusion peak, an important turning point in the life cycle of an innovation, for values of the maximum network degree n of the order of 10^2, which correspond to medium-size real networks with $N \simeq 10^4$ nodes. Then we compare these diffusion times with those of networks with different kinds of degree correlations.

In Section 2, as a preliminary to the analysis of diffusion, we provide the mathematical expressions of the average nearest neighbor degree function $k_{nn}(k)$ and the Newman assortativity coefficient r [14] of BA networks.

In Section 3 we write the network Bass equations and give the diffusion times found from the numerical solutions. We compare these times with those for uncorrelated scale-free networks with exponent $\gamma = 3$ and with those for assortative networks whose correlation matrices are mathematically constructed and studied in another work. Then we apply a method proposed by Newman in [15] to build disassortative correlation matrices and evaluate the corresponding diffusion times. Our results are in qualitative agreement with those by D'Agostino et al. [8] for the SIS model. The minimum diffusion time is obtained for the BA networks, which have, with $n \simeq 10^2$ (corresponding to $N \simeq 10^4$ nodes, see Section 4), $r \simeq -0.10$. The minimum value of r obtained for the disassortative networks (with $\gamma = 3$) is $r \simeq -0.09$, also not far from the values obtained in [8]. For assortative networks, on the contrary, values of r much closer to the maximum $r = 1$ are easily obtained.

Section 4 contains a discussion of the results obtained for the r coefficient and the k_{nn} function of BA networks.

In Section 5 we further reexamine the validity of the mean-field approximation, also by comparison to a version of the Bass model based on the adjacency matrix of BA networks.

In Section 6 we compute diffusion times with heterogeneous p coefficients ("trickle-down" modified Bass model).

In summary, our research objectives in this work have been the following: (a) Analyse the assortativity properties of BA networks, using the correlation functions recently published. These properties are deduced from the Newman coefficient r and from the average nearest neighbor degree function $k_{nn}(k)$. (b) Compute the peak diffusion times of the network Bass model in the mean-field approximation on BA networks, and compare these times with those obtained for other kinds of networks. (c) Briefly discuss the validity of the mean-field approximation, compared to a first-moment closure of the Bass model on BA networks described through the adjacency matrix.

Section 7 contains our conclusions and outlook.

2. The Average Nearest Neighbor Degree Function of Barabasi-Albert (BA) Networks

We emphasize that for the networks considered in this paper a maximal value $n \in \mathrm{N}$ is supposed to exist for the admissible number of links emanating from each node. Notice that, as shown later in the paper (see Section 4.1, 4th paragraph), this is definitely compatible with a high number N of nodes in the network (e.g., with a number N differing from n by various magnitude orders).

Together with the degree distribution $P(k)$ which expresses the probability that a randomly chosen node has k links, other important statistical quantities providing information on the network structure are the degree correlations $P(h \mid k)$. Each coefficient $P(h \mid k)$ expresses the conditional probability that a node with k links is connected to one with h links. In particular, an increasing character of the average nearest neighbor degree function

$$k_{nn}(k) = \sum_{h=1}^{n} hP(h \mid k) \tag{1}$$

is a hallmark of assortative networks (i.e., of networks in which high degree nodes tend to be linked to other high degree nodes), whereas a decreasing character of this function is to be associated with disassortative networks (networks in which high degree nodes tend to be linked to low degree nodes). We recall that the $P(k)$ and $P(h \mid k)$ must satisfy, besides the positivity requirements,

$$\begin{gathered} P(k) \geq 0 \\ \text{and } P(h \mid k) \geq 0, \end{gathered} \tag{2}$$

both the normalizations

$$\begin{gathered} \sum_{k=1}^{n} P(k) = 1 \\ \text{and } \sum_{h=1}^{n} P(h \mid k) = 1, \end{gathered} \tag{3}$$

and the Network Closure Condition (NCC)

$$\begin{aligned} hP(k \mid h)\, P(h) &= kP(h \mid k)\, P(k) \\ &\forall h, k = i = 1, \ldots, n. \end{aligned} \tag{4}$$

A different tool usually employed to investigate structural properties of networks is the Newman assortativity coefficient r also known as the Pearson correlation coefficient [14]. To define it, we need to introduce the quantities e_{jk} which express the probability that a randomly chosen edge links nodes with *excess degree* j and k. Here, the excess degree of a node is meant as its total degree minus one, namely, as the number of all edges emanating from the node except the one under consideration. The distribution of the excess degrees is easily found to be given [14] by

$$q_k = \frac{(k+1)}{\sum_{j=1}^{n} jP(j)} P(k+1). \tag{5}$$

The assortativity coefficient r is then defined as

$$r = \frac{1}{\sigma_q^2} \sum_{k,h=0}^{n-1} kh\left(e_{kh} - q_k q_h\right), \tag{6}$$

where σ_q denotes the standard deviation of the distribution $q(k)$, i.e.,

$$\sigma_q^2 = \sum_{k=1}^{n} k^2 q_k - \left(\sum_{k=1}^{n} k q_k\right)^2. \tag{7}$$

The coefficient r takes values in $[-1, 1]$ and it owes its name to the fact that if $r < 0$, the network is disassortative, if $r = 0$, the network is neutral, and if $r > 0$, the network is assortative. A formula expressing the e_{kh} in terms of known quantities is necessary if one wants to calculate r. This can be obtained as discussed next. Let us move from the elements E_{kh} expressing the number of edges which link nodes with degree k and h, with the only exception that edges linking nodes with the same degree have to be counted twice [16, 17]. Define now $\widetilde{e}_{kh} = E_{kh}/(\sum_{k,h} E_{kh})$. Each $\widetilde{e}_{kh}$ corresponds then to the fraction of edges linking nodes with degree k and h (with the mentioned interpretation of the $\widetilde{e}_{kk}$). We also observe that $e_{k,h} = \widetilde{e}_{k+1,h+1}$ holds true. The degree correlations can be related to the $\widetilde{e}_{kh}$ through the formula

$$P(h \mid k) = \frac{\widetilde{e}_{kh}}{\sum_{j=1}^{n} \widetilde{e}_{kj}} \quad \forall h, k = i = 1, \ldots, n. \tag{8}$$

What is of interest for us here is the following "inverse" formula:

$$\widetilde{e}_{hk} = \frac{P(h \mid k)\, kP(k)}{\sum_{j=1}^{n} jP(j)}. \tag{9}$$

In the rest of this section we discuss the quantities which allow us to calculate the average nearest neighbor degree function $k_{nn}(k)$ and the coefficient r for finite Barabasi-Albert (BA) networks.

The degree distribution of the Barabasi-Albert networks is known to be given by

$$P(k) = \frac{2\beta(\beta+1)}{k(k+1)(k+2)}, \tag{10}$$

where $\beta \geq 1$ is the parameter in the preferential attachment procedure characterizing them [9, 17]. In particular, (10) yields $P(k) \sim c/k^3$ with a suitable constant c for large k.

An explicit expression for the degree correlations $P(h \mid k)$ was given by Fotouhi and Rabbat in [10]. They showed that, for a growing network in the asymptotic limit as $t \longrightarrow \infty$,

$$P(h \mid k) = \frac{\beta}{kh}\left(\frac{k+2}{h+1} - B_{\beta+1}^{2\beta+2}\frac{B_{h-\beta}^{k+h-2\beta}}{B_h^{k+h+2}}\right), \tag{11}$$

with B_j^m denoting the binomial coefficient

$$B_j^m = \frac{m!}{j!(m-j)!}. \tag{12}$$

Since the networks we consider have a maximal number of links n [18], we must normalize the matrix with elements $P(h \mid k)$. We calculate $C_k = \sum_{h=1}^{n} P(h \mid k)$ and take as a new degree correlation matrix the matrix whose (h,k)-element is

$$P_n(h \mid k) = \frac{P(h \mid k)}{C_k}, \tag{13}$$

with $P(h \mid k)$ as in (11).

The average nearest neighbor degree function $k_{nn}(k)$ and the coefficient r can be now easily calculated with software like `Mathematica`, by using (1), (6), (10), and (13). Results are reported and discussed in Section 4.

3. The Bass Diffusion Equation on Complex Networks

In [12, 13] we have reformulated the well-known Bass equation of innovation diffusion

$$\frac{dF(t)}{dt} = [1 - F(t)]\,[p + qF(t)] \tag{14}$$

(where $F(t)$ is the cumulative adopter fraction at the time t and p and q are the innovation and the imitation coefficient, respectively), providing versions suitable for the case in which the innovation diffusion process occurs on a network. The model can be expressed in such a case by a system of n ordinary differential equations:

$$\frac{dG_i(t)}{dt} = [1 - G_i(t)]\left[p + iq\sum_{h=1}^{n} P(h \mid i)\, G_h(t)\right] \qquad i = 1, \ldots, n. \tag{15}$$

The quantity $G_i(t)$ in (15) represents for any $i = 1, \ldots n$ the fraction of potential adopters with i links that at the time t have adopted the innovation. More precisely, denoting by $F_i(t)$ the fraction of the total population composed by individuals with i links, who at the time t have adopted the innovation, we set $G_i(t) = F_i(t)/P(i)$. Further heterogeneity can be introduced allowing also the innovation coefficient (sometimes called publicity coefficient) to be dependent on i. In this case, the equations take the form

$$\frac{dG_i(t)}{dt} = [1 - G_i(t)]\left[p_i + iq\sum_{h=1}^{n} P(h \mid i)\, G_h(t)\right] \qquad i = 1, \ldots, n, \tag{16}$$

and, for example, the p_i can be chosen to be inversely proportional to $P(i)$ or to have a linear dependence, decreasing in i; in the first case more publicity is delivered to the hubs, with ensuing "trickle-down" diffusion, while in the second case a "trickle-up" diffusion from the periphery of the network can be simulated.

The function $f_i(t) = \dot{F}_i(t)$ gives the fraction of new adoptions per unit time in the "link class i" (or "degree class i") i.e., in the subset of individuals having i links. The left panel in Figure 1 shows an example of a numerical solution with plots of all the f_i's, in a case where for graphical reasons we have taken n small ($n = 15$). The underlying network is a BA with $\beta = 1$. As is clear from the plot, the largest fraction of new adopters belongs at all times to the link class with $i = 1$, which reaches its adoption peak later than the others. In general, the more connected individuals are, the earlier they adopt. This phenomenon is quite intuitive and has been evidenced in previous works on mean-field epidemic models; see, for instance, [19]. As discussed in our paper [13], in applications of the Bass model to marketing this may allow to estimate the q coefficient, when it is not known in advance, by monitoring the early occurrence of adoption in the most connected classes. The right panel in Figure 1 shows the total adoption rate $f(t)$ corresponding to the same case as in the left panel. The simple Bass curve (homogeneous model, without underlying network) is also shown for comparison.

In the Bass model, unlike in other epidemic models where infected individuals can return to the susceptible state, the diffusion process always reaches all the population. The function $f(t) = \sum_{i=1}^{n} \dot{F}_i(t)$, which represents the total number of new adoptions per unit time, usually has a peak, as we have seen in the previous example. We choose the time of this peak as a measure of the diffusion time; it is computed for each numerical solution of the diffusion equations by sampling the function $f(t)$. For fixed coefficients of publicity p and imitation q, the time depends on the features of the network. In this paper we consider only scale-free networks with $\gamma = 3$, for comparison with BA networks.

Figure 2 shows the peak times obtained for different networks with maximum degree $n = 100$, as a function of the imitation parameter q. This value of n has been chosen because it corresponds to a number N of nodes of the order of 10^4; this allows a comparison with the results of D'Agostino et al. (see below) and displays finite-size effects, as discussed

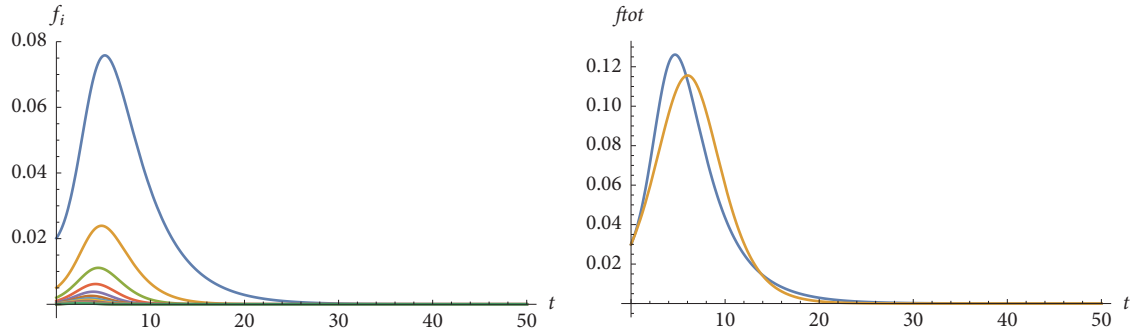


FIGURE 1: *Left panel*: fraction f_i of new adoptions per unit time in the "link class i" (the subset of all individuals having i links), as a function of time, in the Bass innovation diffusion model on a BA network. The parameter β of the network (number of child nodes in the preferential attachment scheme) is $\beta = 1$. The maximum number of links in this example is $n = 15$, while in the comprehensive numerical solutions, whose results are summarized in Figure 2, it is $n = 100$. The adoption peak occurs later for the least connected (and most populated) class with $i = 1$. *Right panel*: cumulative new adoptions $f_{tot} = \sum_{i=1}^{n} f_i$ per unit time, as a function of time, compared with the same quantity for the homogeneous Bass model without an underlying network. The peak of the homogeneous Bass model is slightly lower and shifted to the right. For the values of the model parameters p and q and the measuring unit of time, see Figure 2.

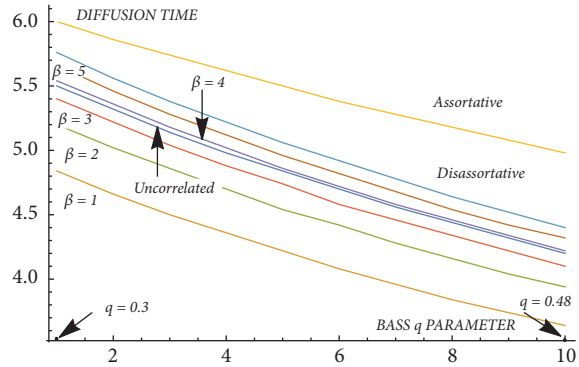


FIGURE 2: Time of the diffusion peak (in years) for the Bass model on different kinds of scale-free networks with exponent $\gamma = 3$, as a function of the imitation coefficient q. All networks have maximum degree $n = 100$ ($N \simeq 10^4$). The q coefficient varies in the range 0.3 – 0.48, corresponding to a typical set of realistic values in innovation diffusion theory [20]. The publicity coefficient is set to the value $p = 0.03$ (see Figure 9 for results with an inhomogeneous p_k depending on the link class $k = 1, \ldots, n$). The lines with $\beta = 1, 2, 3, 4, 5$ correspond to BA networks with those values of β. Their assortativity coefficients are, respectively, $r = -0.104, -0.089, -0.078, -0.071, -0.065$. The disassortative network is built with the Newman recipe (Section 3.4) with $d = 4$ and has $r = -0.084$. The assortative network is built with our recipe (Section 3.3), with $\alpha = 1/2$, and has $r = 0.863$.

in Section 4 (note, however, that such effects are still present with $N \simeq 10^6$ and larger).

It is clear from Figure 2 that diffusion is faster on the BA networks with $\beta = 1, 2, 3$ than on an uncorrelated network. For $\beta = 4$ the diffusion time is almost the same, and for $\beta = 5$ (and $\beta > 5$, not shown) diffusion on the BA network is slower than on the uncorrelated network. On purely disassortative networks diffusion is slightly slower than on an uncorrelated network and much slower than on assortative networks.

3.1. Comparison with the SIS Model. In [8] D'Agostino et al. have studied the dependence of the epidemic threshold and diffusion time for the SIS epidemic model on the assortative or disassortative character of the underlying network. Both the epidemic threshold and the diffusion time were evaluated from the eigenvalues of the adjacency matrix. The networks employed had a number of nodes $N = 10^4$, a scale-free degree distribution with exponent $\gamma = 3$, and were obtained from a BA seed network through a Monte Carlo rewiring procedure which preserves the degree distribution but changes the degree correlations. The Monte Carlo algorithm employs an "Hamiltonian" function which is related to the Newman correlation coefficient r. The values of r explored in this way lie in the range from -0.15 to 0.5 and are therefore comparable with those obtained for our assortative and disassortative matrices.

Although the epidemic model considered and the definition of the diffusion time adopted in [8] are different from ours, there is a qualitative agreement in the conclusions: the diffusion time increases with the assortativity of the network and is at a minimum for values of r approximately equal to -0.10. This value of r corresponds to those of finite BA networks with $\beta = 1$ and $N \simeq 10^4$ and is slightly smaller than the minimum value of r obtained for disassortative networks with a matrix e_{jk} built according to Newman's

recipe (Section 3.4). Note that for those networks the function $k_{nn}(k)$ is decreasing for any k, while for BA networks it is decreasing at small values of k and increasing at large k (Section 4); nevertheless, their r coefficient never becomes significantly less than $\simeq -0.1$, even with other choices in the recipe, as long as $\gamma = 3$. We also recall that the clustering coefficient of the BA networks is exactly zero for $\beta = 1$; for an analysis of the role of the clustering coefficient in epidemic spreading on networks, see, for instance, [21].

A rewiring procedure comparable to that of [8] (without evaluation of the epidemic threshold and diffusion time) has been mathematically described by Van Mieghem et al. [22].

In the next subsections we provide information on the degree correlation matrices and other features of the networks used above for comparison with the BA networks. To compare diffusion peak times on networks which, although characterized by different assortativity/disassortativity properties, share some similarity, we consider networks whose degree distribution $P(k)$ obeys a power-law with exponent three, i.e., is of the form

$$P(k) = \frac{c}{k^3}, \tag{17}$$

where c is the normalization constant.

3.2. Uncorrelated Networks. Let us start with uncorrelated networks. As their name suggests, in these networks the degree correlations $P(h \mid k)$ do not depend on k. They can be easily seen to be given by

$$P(h \mid k) = \frac{hP(h)}{\langle h \rangle}, \tag{18}$$

with $\langle h \rangle = \sum_{h=1}^{n} hP(h)$, and hence their average nearest neighbor degree function (1) is

$$k_{nn}(k) = \frac{\langle h^2 \rangle}{\langle h \rangle}, \tag{19}$$

a constant. The coefficient r is trivially found to be equal to zero.

3.3. A Family of Assortative Networks. We consider now a family of assortative networks we have introduced in [12]. To give here the expressions of their degree correlations $P(h \mid k)$, we need to recall their construction. We start defining the elements of a $n \times n$ matrix P_0 as

$$\begin{aligned} P_0(h \mid k) &= |h-k|^{-\alpha} \quad \text{if } h < k \\ \text{and } P_0(h \mid k) &= 1 \quad \text{if } h = k, \end{aligned} \tag{20}$$

for some parameter $\alpha > 0$, and we define the elements $P_0(h \mid k)$ with $h > k$ in such a way that formula (4) is satisfied by the $P_0(h \mid k)$. Hence, since the normalization $\sum_{h=1}^{n} P_0(h \mid k) = 1$ has to hold true, we compute for any $k = 1, \ldots, n$ the sum $C_k = \sum_{h=1}^{n} P_0(h \mid k)$ and call $C_{max} = \max_{k=1,\ldots,n} C_k$. Then, we introduce a new matrix P_1, requiring that its diagonal elements be given by $P_1(k \mid k) = C_{max} - C_k$ for any $k = 1, \ldots, n$, whereas the nondiagonal elements are the same as those of the matrix P_0: $P_1(h \mid k) = P_0(h \mid k)$ for $h \neq k$. For any $k = 1, \ldots, n$ the column sum $\sum_{h=1}^{n} P_1(h \mid k)$ is then equal to $C_k - 1 + C_{max} - C_k = C_{max} - 1$. Finally, we normalize the entire matrix by setting

$$P(h \mid k) = \frac{1}{(C_{max} - 1)} P_1(h \mid k) \quad \text{for } h, k = 1, \ldots, n. \tag{21}$$

Again, the average nearest neighbor degree function $k_{nn}(k)$ and the coefficient r can be calculated with a software. The increasing character of $k_{nn}(k)$ for a network of the family in this subsection with $\alpha = 1/2$ and $n = 101$ is shown for example in Figure 3.

3.4. A Family of Disassortative Networks. Different models of disassortative networks can be constructed based on a suggestion by Newman contained in [15]. According to it, one can set $e_{kh} = q_k x_h + x_k q_h - x_k x_h$ for $h, k = 0, \ldots, n-1$, where q_k is the distribution of the excess degrees and x_k is any distribution satisfying $\sum_{k=0}^{n-1} x_k = 1$, and with x_k decaying faster than q_k. Choosing, to fix ideas, $x_k = (k+1)^{-\gamma} / \sum_{j=0}^{n-1} (j+1)^{-\gamma}$ with a parameter $\gamma > 2$, we denote $S = \sum_{j=0}^{n-1} (j+1)^{-\gamma}$ and $T = \sum_{j=0}^{n-1} (j+1)^{-2}$ and then set for all $h, k = 0, \ldots, n-1$,

$$\begin{aligned} e_{kh} = \Bigg(& \frac{1}{ST} \left((k+1)^{-2} (h+1)^{-\gamma} + (h+1)^{-2} (k+1)^{-\gamma} \right) \\ & - \frac{1}{S^2} (k+1)^{-\gamma} (h+1)^{-\gamma} \Bigg). \end{aligned} \tag{22}$$

We show in Appendix that the inequalities $0 \leq e_{kh} \leq 1$ hold true for any $h, k = 0, \ldots, n-1$. In view of (8), the degree correlations are then obtained as

$$P(h \mid k) = \frac{e_{k-1,h-1}}{\sum_{j=1}^{n} e_{k-1,j-1}}, \quad \forall h, k = 1, \ldots, n, \tag{23}$$

with the e_{kh} as in (22). It is immediate to check that the coefficient r is negative, see, e.g., [15]. As for the average nearest neighbor degree function $k_{nn}(k)$, it can be calculated with a software. The decreasing character of $k_{nn}(k)$ for a network of the family in this subsection with $\gamma = 4$ and $n = 101$ is shown for example in Figure 4.

4. Discussion

4.1. The Newman Assortativity Coefficient for BA Networks. In [14] Newman reported in a brief table the values of r for some real networks. More data are given in his book [23], Table 8.1. Focusing on disassortative scale-free networks for which the scale-free exponent γ is available, one realizes that their negative r coefficient is generally small in absolute value, especially when γ is equal or close to 3, for instance,

(i) `www.nd.edu`: γ from 2.1 to 2.4, $r = -0.067$

(ii) Internet: $\gamma = 2.5$, $r = -0.189$

(iii) Electronic circuits: $\gamma = 3.0$, $r = -0.154$

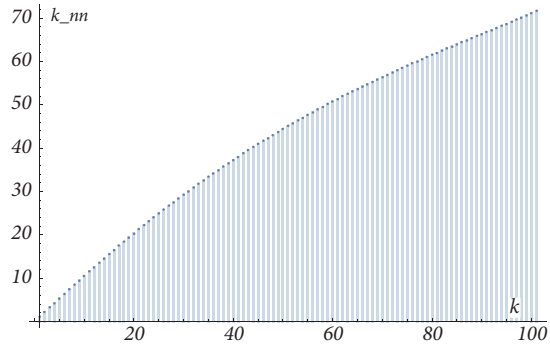


Figure 3: Function k_{nn} for an assortative network as in Section 3.3 with $\alpha = 1/2$, $n = 101$ ($N \simeq 10^4$).

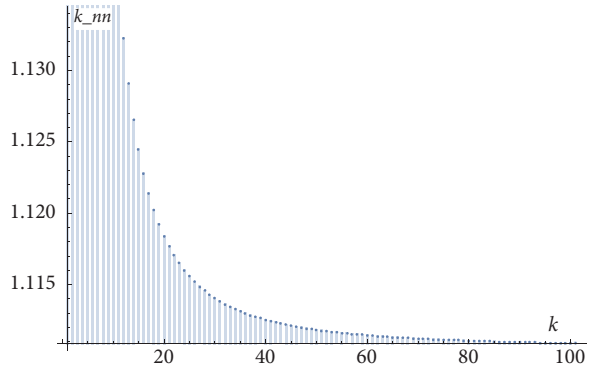


Figure 4: Function k_{nn} for a disassortative network as in Section 3.4 with $\gamma = 4$, $n = 101$ ($N \simeq 10^4$).

The size of these networks varies, being of the magnitude order of $N \approx 10^4$ to $N \approx 10^6$. The data mentioned above have been probably updated and extended, but in general it seems that scale-free disassortative networks with these values of γ tend to have an r coefficient that is much closer to 0 than to -1. As we have seen, this also happens with disassortative scale-free networks mathematically defined, whose degree correlations are given by a procedure also introduced by Newman.

For ideal BA networks, Newman gave in [14] an asymptotic estimate of the r coefficient based on the correlations computed in [24] and concluded that r is negative but vanishes as $(\log N)^2/N$ in the large-N limit. Other authors share the view that BA networks are essentially uncorrelated [19, 25]. The smallness of their r coefficient is also confirmed by numerical simulations, in which the network is grown according to the preferential attachment scheme, even though the results of such simulations are affected by sizable statistical errors when N varies between $\approx 10^4$ and $\approx 10^6$.

In the recent paper [26] Fotouhi and Rabbat use their own expressions for the $p(l,k)$ correlations (expressing the fraction of links whose incident nodes have degrees l and k) to compute the asymptotic behavior of r according to an alternative expression given by Dorogovtsev and Mendes [27]. They conclude that the estimate $r \simeq (\log N)^2/N$ given by Newman is not correct. They find $|r| \approx (\log n)^2/n$. In order to relate the network size N to the largest degree n, they use the relation $n \simeq \sqrt{N}$ based on the continuum-like criterium $\int_n^\infty P(k)dk = N^{-1}$ [28]. So in the end $|r| \approx (\log N)^2/\sqrt{N}$. They check that this is indeed the correct behavior by performing new large-scale simulations.

We have computed r using the original definition by Newman [14], based on the matrix e_{jk}, and the exact $P(h \mid k)$ coefficients, with n up to 15000, which corresponds to $N \simeq 225000000$ (Figure 5). We found a good agreement with the estimate of Fotouhi and Rabbat. Note, however, that although the "thermodynamic" limit of r for infinite N is zero, for finite N the value of r cannot be regarded as vanishingly small, in consideration of what we have seen above for disassortative scale-free networks with $\gamma = 3$.

4.2. The Function $k_{nn}(k)$. An early estimate of the function $k_{nn}(k)$ valid also for BA networks has been given by Vespignani and Pastor-Satorras in an Appendix of their book on the structure of the Internet [29]. Their formula is based in turn on estimates of the behavior of the conditional probability $P(k' \mid k)$ which hold for a growing scale-free network with a degree distribution $P(k) \propto k^{-\gamma}$. This formula reads

$$P\left(k' \mid k\right) \propto k'^{-(\gamma-1)} k^{-(3-\gamma)} \tag{24}$$

and holds under the condition $1 \ll k' \ll k$. Using the exact $P(k' \mid k)$ coefficients of Fotouhi and Rabbat it is possible to check the approximate validity of this formula.

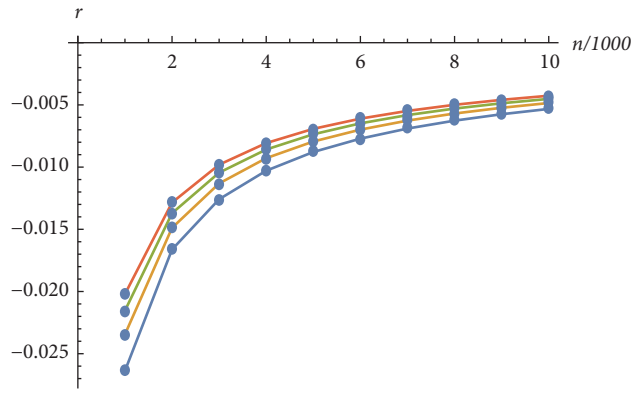


FIGURE 5: Newman assortativity coefficient r for BA networks with $\beta = 1, 2, 3, 4$ and largest degree $n = 1000, 2000, \ldots, 10000$. The upper curve is for $\beta = 1$, and the curves for the values $\beta = 2, 3, 4$ follow, from the top to the bottom.

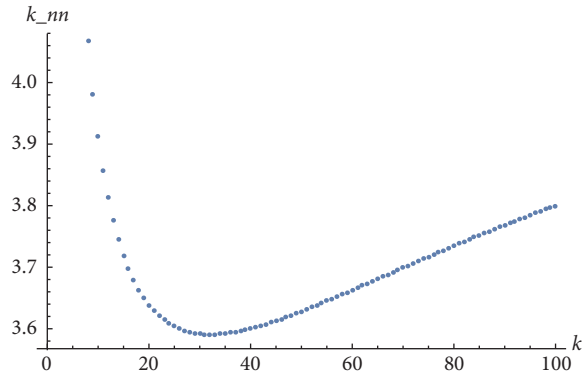


FIGURE 6: Function k_{nn} for a BA network with $\beta = 1$, $n = 100$ (largest degree; the corresponding number of nodes is $N \simeq 10^4$).

In [29] the expression (24) is then used to estimate $k_{nn}(k)$. Since, for $\gamma = 3$, $P(k' \mid k)$ does not depend on k, the conclusion is that k_{nn} is also independent from k. It is not entirely clear, however, how the condition $1 \ll k' \ll k$ can be fulfilled when the sum over k' is performed, and how the diverging factor $\sum_{k'=1}^{n}(1/k') \simeq \ln(n) \simeq (1/2)\ln(N)$ should be treated.

In their recent work [26], Fotouhi and Rabbat use their own results for the $P(k' \mid k)$ coefficients in order to estimate $k_{nn}(k)$ in the limit of large k. They find $k_{nn}(k) \approx \beta \ln(n)$ and cite a previous work [30] which gives the same result, in the form $k_{nn}(k) \approx (1/2)\beta \ln(N)$ (we recall that $n \simeq \sqrt{N}$). In this estimate we can observe, as compared to [29], the explicit presence of the parameter β and the diverging factor $\ln(n)$.

Concerning the Newman assortativity coefficient r, Fotouhi and Rabbat also make clear that even though $r \longrightarrow 0$ when $N \longrightarrow \infty$, this does not imply that the BA networks are uncorrelated, and in fact the relation $k_{nn}(k) = \langle k^2 \rangle / \langle k \rangle$, valid for uncorrelated networks, does not apply to BA networks.

A direct numerical evaluation for finite n of the function $k_{nn}(k)$ based on the exact $P(k' \mid k)$ shows further interesting features. As can be seen in Figures 6 and 7, the function is decreasing at small k and slightly and almost linearly increasing at large k. Note that this happens for networks of medium size ($n = 100$, $N \simeq 10^4$) like those employed in our numerical solution of the Bass diffusion equations and employed in [8], but also for larger networks (for instance, with $n = 1000$, $N \simeq 10^6$, compare Figure 7). It seems that the "periphery" of the network, made of the least connected nodes, has a markedly disassortative character, while the hubs are slightly assortative. (On general grounds one would instead predict for finite scale-free networks a small structural disassortativity at large k [17].)

Since evidence on epidemic diffusion obtained so far indicates that generally the assortative character of a network lowers the epidemic threshold and the disassortative character tends to make diffusion faster once it has started, this "mixed" character of the finite BA networks appears to facilitate spreading phenomena and is consistent with our data on diffusion time (Section 3). Note in this connection that some real networks also turn out to be both assortative and disassortative, in different ranges of the degree k; compare the examples in [17], Ch. 7.

Finally, we would like to relate our numerical findings for the function k_{nn} to the general property (compare, for instance, [5])

$$\left\langle k^2 \right\rangle = \sum_{k=1}^{n} kP(k) K(k, n), \tag{25}$$

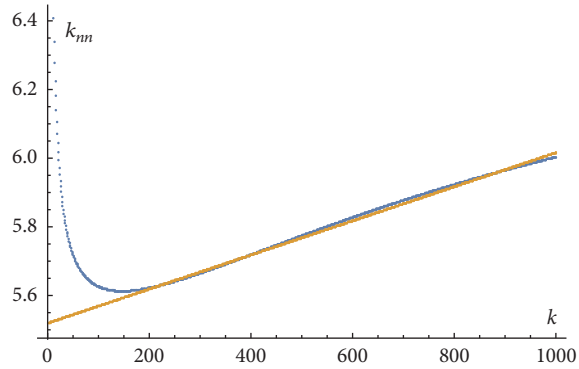


Figure 7: Function k_{nn} for a BA network with $\beta = 1$, $n = 1000$ ($N \simeq 10^6$). A linear fit for large k is also shown. The slope is approximately equal to $5 \cdot 10^{-4}$; a comparison with Figure 6 shows that the slope decreases with increasing n, but is still not negligible.

where for simplicity K denotes the function k_{nn} and the dependence of this function on the maximum degree n is explicitly shown. For BA networks with $\beta = 1$ we have at large n on the l.h.s. of (25), from the definition of $\langle k^2 \rangle$,

$$\left\langle k^2 \right\rangle = \sum_{k=1}^{n} k^2 \frac{4}{k\,(k+1)\,(k+2)} = 4 \ln (n) + o(n), \qquad (26)$$

where the symbol $o(n)$ denotes terms which are constant or do not diverge in n.

For the expression on the r.h.s. of (25) we obtain

$$\sum_{k=1}^{n} kP(k)\,K(k,n) = \sum_{k=1}^{n} \frac{4}{(k+1)\,(k+2)} K(k,n) \qquad (27)$$

Equations (26) and (27) are compatible, in the sense that their diverging part in n is the same, in two cases: (1) if for large k we have $K(k,n) \simeq a \ln(n)$, where a is a constant; this is true because in that case the sum on k is convergent; (2) if more generally $K(k,n) \simeq a \ln(n) + b(n)k$; this is still true because also for the term $b(n)k$ the sum in k leads to a result proportional to $\ln(n)$. Case (2) appears to be what happens, according to Figures 6 and 7.

5. Validity of the Mean-Field Approximation

The validity of the heterogeneous mean-field approximation in epidemic diffusion models on networks has been discussed in [11, 31] and references. As far as the form of the equation is concerned, the Bass model is very similar to other epidemic models. A network which is completely characterized by the functions $P(k)$, $P(h \mid k)$, is called in [31] a "Markovian network". Mean-field statistical mechanics on these networks has been treated in [32]. Recent applications can be found, for instance, in [33]. It has, in a certain sense, an axiomatic value and it is well defined in the statistical limit of infinite networks. Also the BA networks studied by Fotouhi and Rabbat are defined in this limit. In our work we are mainly interested into the properties of the networks, therefore when we speak of their properties with respect to diffusion, we make reference to this kind of idealized diffusion. Note that the mean-field approximation employed has a quite rich structure (whence the denomination "heterogeneous"), since the population can be divided into a large number of connectivity classes, and the functions $P(k)$, $P(h \mid k)$ contain a lot of networking information about these classes.

It may therefore be of interest to check at least numerically if the mean-field approximation is a good approximation of the full Bass model in this case. For this purpose one may consider a network formulation of the model which employs the adjacency matrix $\{a_{ij}\}$. The coupled differential equations for a network with N nodes are in this formulation

$$\frac{dF_i(t)}{dt} = \left[1 - F_i(t)\right] \left(p + q \sum_{j=1}^{N} a_{ij} F_j(t) \right), \qquad i = 1, \ldots, N. \qquad (28)$$

The solution $F_i(t)$ defines, for each node, an adoption level which grows as a function of time from $F_i = 0$ at $t = 0$, to $F_i = 1$ when t goes to infinity. The "bare" q parameter is renormalized with the average connectivity $\langle k \rangle$. This model can be seen as a first-moment closure of a stochastic model [23] and is actually more suitable to describe innovation diffusion among firms or organizations, under the assumption that inside each organization innovations diffuse in a gradual way. In fact, in forthcoming work we apply the model to real data concerning networks of enterprises. It is also useful, however, in order to illustrate the difference between a mean-field approach and a completely realistic approach. For the case of a BA network it is possible to generate the adjacency matrix with a random preferential attachment algorithm. This matrix is then fed to a code which solves the differential equations (28) and computes for each node the characteristic diffusion times. The results clearly show that in general different nodes with the same degree adopt at different rates (unlike in the mean-field approximation). It happens, for instance, as intuitively expected, that a node with degree 1 which is connected to a big central hub adopts more quickly than a node with the same degree that has been attached to the network periphery in the final phase of the growth. However, these fluctuations can be estimated numerically, and it turns out that mean-field approximations for the diffusion times agree with the average times within

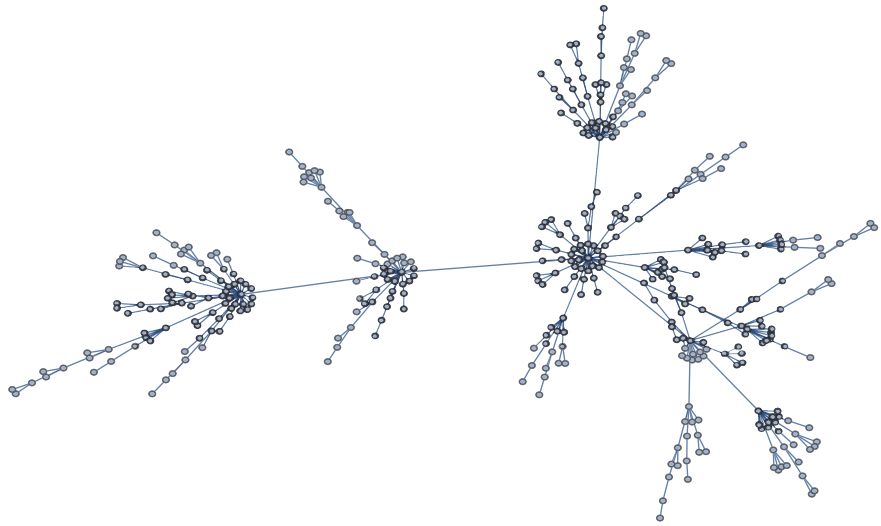


Figure 8: An example of a BA network with 400 nodes and $\beta = 1$, to which (28) has been applied for comparison to the corresponding mean-field approximation.

standard errors. For instance, in a realization of a BA network with $\beta = 1$, $N = 400$ (Figure 8), we obtain (considering the most numerous nodes, for which statistics is reliable) the following:

Degree 1: 266 nodes, with average diffusion peak time 5.2, $\sigma = 1.3$

Degree 2: 67 nodes, with average diffusion peak time 4.8, $\sigma = 1.1$

Degree 3: 33 nodes, with average diffusion peak time 4.3, $\sigma = 0.9$

With the mean-field model, with the same bare $p = 0.03$ and $q = 0.35$, we obtain the following:

Degree-1 nodes: diffusion peak time 5.2

Degree-2 nodes: diffusion peak time 4.9

Degree-3 nodes: diffusion peak time 4.6

The agreement is reasonable, also in consideration of the relatively small value of N. Concerning this, a final remark is in order: while random generation of BA networks mainly yields networks with maximum degree close to the most probable value $n = \sqrt{N}$, sizable fluctuations may occur as well. The network in Figure 8, for instance, has $N = 400$ and a hub with degree 35. We intentionally included this network in our checks, in order to make sure that the mean-field approximation can handle also these cases. This turns out to be true, provided the *real* value of the maximum degree is used (the mean-field peak times given above are obtained with $n = 35$). In fact, the full numerical solutions of (28) routinely show, as expected, that for fixed values of N the presence of large hubs speeds up the overall adoption. For an "average" network, the mean-field approximation works well just with $n = \sqrt{N}$.

6. Diffusion Times with Heterogeneous p Coefficients

In the Bass model the publicity coefficient p gives the adoption probability per unit time of an individual who has not yet adopted the innovation, independently from the fraction of current adopters. Therefore, it is not due to the word-of-mouth effect, but rather to an external stimulus (advertising) which is received in the same way by all individuals. The intensity of this stimulus is in turn proportional to the advertising expenses of the producer or seller of the innovation. One can therefore imagine the following alternative to the uniform dissemination of ads to all the population: the producer or seller invests for advertising to each individual in inverse proportion to the population of the individual's link class, so as to speed up adoption in the most connected classes and keep the total expenses unchanged. In terms of the p_i coefficients in (16) this implies $p_i \propto 1/P(i)$, with normalization $\sum_{i=1}^{n} p_i P(i) = p$ [12].

As can be seen also from Figure 9, this has the effect of making the total diffusion faster or slower, depending on the kind of network. The differences observed in the presence of a homogeneous p coefficient (Figure 2) are now amplified. We observe that diffusion on BA networks is now always faster than on uncorrelated networks, independently from β. It is also remarkable how slow diffusion becomes on assortative networks in this case. A possible interpretation is the following: due to the low epidemic threshold of assortative networks, the targeted advertising on the hubs (which are well connected to each other) causes them to adopt very quickly, but this is not followed by easy diffusion on the whole network. The BA networks appear instead to take advantage, in the presence of advertising targeted on the hubs, both from their low threshold and from a stronger linking to the periphery.

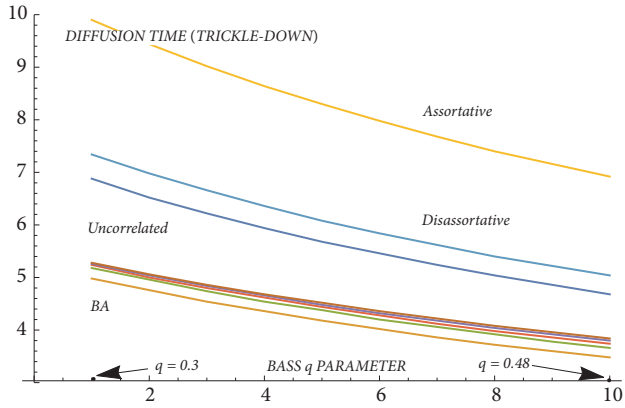


FIGURE 9: Time of the diffusion peak (in years) for the "trickle-down" modified Bass model on different kinds of scale-free networks with exponent $\gamma = 3$, as a function of the imitation coefficient q. All networks have maximum degree $n = 100$. The q coefficient varies in the range $0.3 - 0.48$, corresponding to a typical set of realistic values in innovation diffusion theory [20]. The publicity coefficient p_k of the link class k, with $k = 1, \ldots, n$, varies in inverse proportion to $P(k)$ (Section 6). The low-lying group of lines corresponds to BA networks with $\beta = 1, 2, 3, 4, 5$ and the same assortativity coefficients as in Figure 2. The disassortative network is built with the Newman recipe (Section 3.4) with $d = 4$ and has $r = -0.084$. The assortative network is built with our recipe (Section 3.3), with $\alpha = 1/4$, and has $r = 0.827$.

7. Conclusions

In this work we have studied the assortativity properties of BA networks with maximum degree n which is finite, but large (up to $n \simeq 10^4$, corresponding to a number of nodes $N \simeq 10^8$). These properties were not known in detail until recently; a new decisive input has come from the exact calculation of the conditional probabilities $P(h \mid k)$ by Fotouhi and Rabbat [10]. We have done an explicit numerical evaluation of the average nearest neighbor degree function $k_{nn}(k)$, whose behavior turns out to be peculiar and unexpected, exhibiting a coexistence of assortative and disassortative correlations, for different intervals of the node degree k. These results have been compared with previous estimates, both concerning the function $k_{nn}(k)$ and the Newman assortativity index r.

The next step has been to write the Bass innovation diffusion model on BA networks, following the mean-field scheme we have recently introduced and tested on generic scale-free networks. This allows computing numerically the dependence of the diffusion peak time (for $n \simeq 10^2$) from the model's parameters and especially from the features of the network. We have thus compared the diffusion times on BA networks with those on uncorrelated, assortative and disassortative networks (the latter built, respectively, with our mathematical recipes (20) and (21) and with Newmans's recipe (22)).

The BA networks with small values of β (β is the number of child nodes in the preferential attachment scheme) turn out to have the shortest diffusion time, probably due to their mixed assortative/disassortative character: diffusion appears to start quite easily among the (slightly assortative) hubs and then to proceed quickly in the (disassortative) periphery of the network. This interpretation is confirmed by the fact that, in a modified "trickle-down" version of the model with enhanced publicity on the hubs, the anticipation effect of BA networks compared to the others is stronger and almost independent from β.

Concerning the dependence of the diffusion time on the values of the r coefficient, we have found a qualitative agreement with previous results by D'Agostino et al. [8]. We stress, however, that the use of two-point correlations in this analysis is entirely new.

In forthcoming work we shall analyse mathematically the construction and the properties of the mentioned families of assortative networks, from which we only have chosen here a few samples for comparison purposes, because the focus in this paper has been on BA networks.

Appendix

We show here that e_{kh} in (22) satisfies $0 \leq e_{kh} \leq 1$ for all $h, k = 0, \ldots, n-1$. If $\gamma = 2 + d$ with $d > 0$, $e_{k-1,h-1}$ can be rewritten for any $h, k = 1, \ldots, n$ as

$$e_{k-1,h-1} = \frac{\left(k^d + h^d\right) \sum_{j=1}^{n} \left(1/j^{\gamma}\right) - \sum_{j=1}^{n} \left(1/j^{2}\right)}{k^{\gamma} h^{\gamma} \sum_{j=1}^{n} \left(1/j^{2}\right) \left(\sum_{j=1}^{n} \left(1/j^{\gamma}\right)\right)^{2}}. \tag{A.1}$$

The nonnegativity of $e_{k-1,h-1}$ is hence equivalent to that of the numerator of (A.1). This is, for all $h, k = 1, \ldots, n$, greater than or equal to

$$2 \sum_{j=1}^{n} \frac{1}{j^{\gamma}} - \sum_{j=1}^{\infty} \frac{1}{j^{2}} \geq 2 - \frac{\pi}{6} > 0. \tag{A.2}$$

To show that $e_{k-1,h-1} \leq 1$, we distinguish two cases:

(i) if $k = 1$ and $h = 1$, the expression on the r.h.s. in (A.1) is equal to

$$\frac{2\sum_{j=1}^{n}\left(1/j^{\gamma}\right)-\sum_{j=1}^{n}\left(1/j^{2}\right)}{\sum_{j=1}^{n}\left(1/j^{2}\right)\left(\sum_{j=1}^{n}\left(1/j^{\gamma}\right)\right)^{2}} \leq \frac{2\sum_{j=1}^{n}\left(1/j^{2}\right)-\sum_{j=1}^{n}\left(1/j^{2}\right)}{\sum_{j=1}^{n}\left(1/j^{2}\right)\left(\sum_{j=1}^{n}\left(1/j^{\gamma}\right)\right)^{2}} \leq \frac{1}{\left(\sum_{j=1}^{n}\left(1/j^{\gamma}\right)\right)^{2}} \leq 1; \tag{A.3}$$

(ii) otherwise, i.e., if at least one among k and h is greater than 1, the expression on the r.h.s. in (A.1) is no greater than

$$\frac{\left(k^{d}+h^{d}\right)\sum_{j=1}^{n}\left(1/j^{2+d}\right)}{k^{2+d}h^{2+d}\sum_{j=1}^{n}\left(1/j^{2}\right)\left(\sum_{j=1}^{n}\left(1/j^{2+d}\right)\right)^{2}} \leq \frac{\left(k^{d}+h^{d}\right)}{k^{d}h^{d}}\frac{1}{k^{2}h^{2}} \leq 1. \tag{A.4}$$

Data Availability

No data were used to support this study.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the Open Access Publishing Fund of the Free University of Bozen-Bolzano.

References

[1] M. J. Keeling and K. T. D. Eames, "Networks and epidemic models," *Journal of The Royal Society Interface*, vol. 2, no. 4, pp. 295–307, 2005.

[2] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, and A. Vespignani, "Epidemic processes in complex networks," *Reviews of Modern Physics*, vol. 87, no. 3, pp. 925–979, 2015.

[3] A. M. Porter and P. J. Gleeson, "Dynamical systems on networks," in *Frontiers in Applied Dynamical Systems: Reviews and Tutorials*, vol. 4, 2016.

[4] R. Pastor-Satorras and A. Vespignani, "Epidemic dynamics and endemic states in complex networks," *Physical Review E*, vol. 63, no. 6, Article ID 066117, 2001.

[5] M. Boguñá, R. Pastor-Satorras, and A. Vespignani, "Absence of Epidemic Threshold in Scale-Free Networks with Degree Correlations," *Physical Review Letters*, vol. 90, no. 2, 2003.

[6] V. M. Eguíluz and K. Klemm, "Epidemic threshold in structured scale-free networks," *Physical Review Letters*, vol. 89, no. 10, Article ID 108701, 2002.

[7] P. Blanchard, C. Chang, and T. Krüger, "Epidemic Thresholds on Scale-Free Graphs: the Interplay between Exponent and Preferential Choice," *Annales Henri Poincaré* , vol. 4, no. S2, pp. 957–970, 2003.

[8] G. D'Agostino, A. Scala, V. Zlatić, and G. Caldarelli, "Robustness and assortativity for diffusion-like processes in scale-free networks," *Europhys. Lett*, vol. 97, no. 6, p. 68006, 2012.

[9] A. Barabasi and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[10] B. Fotouhi and M. G. Rabbat, "Degree correlation in scale-free graphs," *European Physical Journal B*, vol. 86, no. 12, p. 510, 2013.

[11] A. Vespignani, "Modelling dynamical processes in complex socio-technical systems," *Nature Physics*, vol. 8, no. 1, pp. 32–39, 2012.

[12] M. Bertotti, J. Brunner, and G. Modanese, "The Bass diffusion model on networks with correlations and inhomogeneous advertising," *Chaos, Solitons & Fractals*, vol. 90, pp. 55–63, 2016.

[13] M. L. Bertotti, J. Brunner, and G. Modanese, "Innovation diffusion equations on correlated scale-free networks," *Physics Letters A*, vol. 380, no. 33, pp. 2475–2479, 2016.

[14] M. E. J. Newman, "Assortative mixing in networks," *Physical Review Letters*, vol. 89, no. 20, Article ID 208701, 2002.

[15] M. E. Newman, "Mixing patterns in networks," *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics*, vol. 67, no. 2, 2003.

[16] M. E. J. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, no. 2, pp. 167–256, 2003.

[17] A.-L. Barabási, *Network Science*, Cambridge University Press, Cambridge, UK, 2016.

[18] R. Pastor-Satorras and A. Vespignani, "Epidemic dynamics in finite size scale-free networks," *Phys. Rev. E*, vol. 65, no. 3, Article ID 035108, 2002.

[19] M. Barthélemy, A. Barrat, R. Pastor-Satorras, and A. Vespignani, "Dynamical patterns of epidemic outbreaks in complex heterogeneous networks," *Journal of Theoretical Biology*, vol. 235, no. 2, pp. 275–288, 2005.

[20] Z. Jiang, F. M. Bass, and P. I. Bass, "Virtual Bass Model and the left-hand data-truncation bias in diffusion of innovation studies," *International Journal of Research in Marketing*, vol. 23, no. 1, pp. 93–106, 2006.

[21] V. Isham, J. Kaczmarska, and M. Nekovee, "Spread of information and infection on finite random networks," *Physical Review E*, vol. 83, no. 4, Article ID 046128, 2011.

[22] P. V. Mieghem, H. Wang, X. Ge, S. Tang, and F. A. Kuipers, "Influence of assortativity and degree-preserving rewiring on the spectra of networks," *The European Physical Journal B*, vol. 76, no. 4, pp. 643–652, 2010.

[23] M. Newman, *Networks: An Introduction*, Oxford University Press, Oxford, UK, 2010.

[24] P. L. Krapivsky and S. Redner, "Organization of growing random networks," *Physical Review E*, vol. 63, no. 6, Article ID 066123, 2001.

[25] R. Noldus and P. Van Mieghem, "Assortativity in complex networks," *Journal of Complex Networks*, vol. 3, no. 4, pp. 507–542, 2015.

[26] B. FOTOUHI and M. RABBAT, "Temporal evolution of the degree distribution of alters in growing networks," *Network Science*, vol. 6, no. 01, pp. 97–155, 2018.

[27] S. N. Dorogovtsev and J. F. F. Mendes, "Evolution of networks," *Advance in Physics*, vol. 51, no. 4, pp. 1079–1187, 2002.

[28] M. Boguña, R. Pastor-Satorras, and A. Vespignani, "Cut-offs and finite size effects in scale-free networks," *The European Physical Journal B*, vol. 38, no. 2, pp. 205–209, 2004.

[29] R. Pastor-Satorras and A. Vespignani, *Evolution and Structure of the Internet: A Statistical Physics Approach*, Cambridge University Press, Cambridge, UK, 2004.

[30] A. Barrat and R. Pastor-Satorras, "Rate equation approach for correlations in growing network models," *Physical Review E*, vol. 71, no. 3, Article ID 036127, 2005.

[31] M. Boguñá and R. Pastor-Satorras, "Epidemic spreading in correlated complex networks," *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics*, vol. 66, no. 4, Article ID 047104, 2002.

[32] J. Marro and R. Dickman, *Nonequilibrium Phase Transitions in Lattice Models*, Cambridge University Press, Cambridge, UK, 1999.

[33] C. Osorio and C. Wang, "On the analytical approximation of joint aggregate queue-length distributions for traffic networks: A stationary finite capacity Markovian network approach," *Transportation Research Part B: Methodological*, vol. 95, pp. 305–339, 2017.

Hindawi
Complexity
Volume 2019, Article ID 5039097, 13 pages
https://doi.org/10.1155/2019/5039097

Research Article

Prediction of Ammunition Storage Reliability Based on Improved Ant Colony Algorithm and BP Neural Network

Fang Liu,[1] **Hua Gong**,[1] **Ligang Cai**,[2] **and Ke Xu**[1]

[1]*College of Science, Shenyang Ligong University, Shenyang 110159, China*
[2]*College of Science, Shenyang University of Technology, Shenyang 110178, China*

Correspondence should be addressed to Hua Gong; gonghua@sylu.edu.cn

Received 30 November 2018; Accepted 24 February 2019; Published 18 March 2019

Guest Editor: Pedro Palos

Storage reliability is an important index of ammunition product quality. It is the core guarantee for the safe use of ammunition and the completion of tasks. In this paper, we develop a prediction model of ammunition storage reliability in the natural storage state where the main affecting factors of ammunition reliability include temperature, humidity, and storage period. A new improved algorithm based on three-stage ant colony optimization (IACO) and BP neural network algorithm is proposed to predict ammunition failure numbers. The reliability of ammunition storage is obtained indirectly by failure numbers. The improved three-stage pheromone updating strategies solve two problems of ant colony algorithm: local minimum and slow convergence. Aiming at the incompleteness of field data, "zero failure" data pretreatment, "inverted hanging" data pretreatment, normalization of data, and small sample data augmentation are carried out. A homogenization sampling method is proposed to extract training and testing samples. Experimental results show that IACO-BP algorithm has better accuracy and stability in ammunition storage reliability prediction than BP network, PSO-BP, and ACO-BP algorithm.

1. Introduction

Reliability is the primary quality index of military products. On the battlefield, the unreliable ammunition products will lead to the failure of military tasks or endanger the lives of our soldiers. In the weapon system, the reliability is the core of the design, development, and maintenance of ammunition products. Since most of the life cycle of ammunition products is the storage period, the storage reliability of ammunition directly affects the field reliability of ammunition. Quantitative prediction of ammunition products storage reliability is the key link of ammunition reliability engineering.

In recent years, scholars have carried out extensive research on reliability in their respective fields. Many models and methods are devised to predict and evaluate the reliability of different products and systems (see [1–6]). On the reliability of ammunition storage, the traditional methods are based on natural storage condition data and accelerated testing data, respectively. The reliability of ammunition storage is predicted by mathematical statistical methods. Based on natural storage condition data, a Poisson reliability mathematical model is established in [7] to predict ammunition storage life. Binary Logic Regression (BLR) model is obtained in [8] to evaluate recent system failures in non-operational storage. Under the assumption of exponential distribution, a storage reliability prediction model based on time-truncated data is established in [9]. In [10], Gamma distribution is used to fit the temperature distribution in the missile during storage, and proportional risk model is created to predict the effect of temperature on product reliability. In order to forecast the residual storage life, second-order continuous-time homogeneous Markov model and stochastic filtering model are utilized in [11] and [12], respectively. In [13], an E-Bayes statistical model is proposed to predict storage life by using the initial failure number.

Based on the accelerated test data, the storage life of fuze is evaluated in [14] by applying step stress accelerated test. The ammunition storage reliability is predicted in [15] based on ammunition failure mechanism and accelerated life model. In [16], the storage reliability function expression of ammunition is established based on the state monitoring data of accelerated test. The Arrhenius acceleration algorithm

is used in [17] to design a new model to estimate the acceleration factor. In [18], a prediction method combining accelerated degradation test with accelerated life test is proposed. Monte Carlo method is used to generate pseudo-failure life data under various stress levels to evaluate reliability and predict life of products. Since the traditional statistical methods usually need the original life distribution information in solving the reliability prediction problems, they are conservative in solving the uncertain life distribution, highly nonlinear problems, and small sample data problems.

Many scholars recently consider intelligent algorithms in reliability prediction. Specifically, a comprehensive prediction model which combines SVM with Ayers methods is given to predict the reliability of small samples in [19]. The wavelet neural network is established in [20] to predict the data block of hard disk storage system. Two kinds of missile storage reliability prediction models are designed and compared in [21]. Two models are BP network and RBF network. Artificial neural network equipment degradation model is proposed in [22, 23] to predict the remaining life of equipment. A particle swarm optimization model is employed in [24–26] to solve the reliability optimization problem with uncertainties. The intelligent algorithms based on the original data keep the information of data to the greatest extent. It is noted that the intelligent prediction algorithms do not depend on the prior distribution of ammunition life. The exploration process of prior distribution of raw data is eliminated. The intelligent algorithms provide a new way to predict the reliability of ammunition storage. However, there are still many problems need to be solved when applying intelligent algorithms to reliability prediction, such as the local minimum, slow convergence speed, and high sensitivity of SVM to data missing.

In this paper, a prediction model of ammunition failure number is proposed. A new three-stage ACO and BP neural network is created in the model. This prediction model excavates the mathematical relationship between the storage temperature, humidity, and period of ammunition under natural conditions and the number of ammunition failure. The reliability of ammunition can be obtained indirectly by the failure number. In the aspect of sample selection, "zero failure" data pretreatment, "inverted hanging" data pretreatment, normalization of small sample data augmentation, and homogenization sampling are adopted to achieve sample integrity. Experimental results show that our IACO-BP out-performs BP, ACO-BP, and PSO-BP.

The rest of this paper is organized as follows. Section 2 introduces the basic theory of ammunition storage reliability. Section 3 presents IACO-BP algorithm and describes the implementation details. Section 4 demonstrates data pretreatment methods and training sample and testing sample extraction strategy. Section 5 shows the experimental results and compares algorithm performance with BP, PSO-BP, and ACO-BP. The conclusions are denoted in Section 6.

2. Basic Theory of Ammunition Storage Reliability

The storage reliability of ammunition is the ability of ammunition to fulfill the specified functions within the specified storage time and under the storage conditions. The storage reliability of ammunition $R(t)$ is the probability of keeping the specified function. Here, $R(t) = P(\eta > t)$, where η is the storage time before ammunition failure and t is the required storage time.

According to the definition of reliability, we can obtain the following:

$$R(t) = \frac{N_{t=0} - f(t)}{N_{t=0}} \tag{1}$$

where $N_{t=0}$ is the initial ammunition number and $f(t)$ is the cumulative failure number of ammunitions from the initial time to the time t.

The storage of modern ammunitions is a complex system which composes of chemical, mechanical, and photoelectric materials. Ammunitions are easy to be affected by different environmental factors. For example, nonmetallic mold leads to the insulation ability of insulation materials decline, the charge is moisture to reduce the explosive force of ammunition, and plastic hardening results in self-ignition of ammunition. Researches show that temperature and humidity are the main two factors that affect ammunition reliability, and storage period is an important parameter of ammunition reliability.

3. IACO and BP Neural Network

3.1. Traditional Ant Colony Optimization Algorithm (ACO). Ant Colony Optimization (ACO) proposed by Dorigo is a global strategy intelligent optimization algorithm, which simulates the foraging behaviors of ants in [27]. Ants that set out to find food always leave the secretion which is called pheromone on the path. The following ants adaptively decide the shortest path to the food position by the residual pheromone concentration. The overall cooperative behaviors of ant colony foraging constitute a positive feedback phenomenon. Ant colony algorithm simulates the optimization mechanism of information exchange and cooperation among individuals in this process. It searches the global optimal solution through information transmission and accumulation in the solution space.

The main steps of ant colony algorithm are described as follows.

Step 1 (setting the initial parameters value). The number of ants is set to Q. η_i $(1 \le i \le n)$ is n parameters to be optimized. Each parameter has M values in the range of values. These values form set A_{η_i}. At the initial time, each element carries the same concentration pheromone that is recorded as $\xi_j(A_{\eta_i})(0) = a$ $(1 \le j \le M)$. The maximum number of iterations is $N_{\max}$.

Step 2 (choosing paths according to strategy). Start the ants. Each ant starts from set A_{η_i} and chooses the foraging path according to (2) route selection strategy.

$$P\left(\xi_j^k\left(A_{\eta_i}\right)\right) = \frac{\xi_j\left(A_{\eta_i}\right)}{\sum_{j=1}^{M}\xi_j\left(A_{\eta_i}\right)} \tag{2}$$

Equation (2) describes the probability that ant k selects j value of parameter i. Ants determine the next foraging position according to the probability maximum selection rule.

Step 3 (updating pheromone). According to selection strategy in Step 2, each ant selects a location element in each set A_{η_i}. τ is the time of this process. When ants finish a cycle of parameter selection, pheromones are updated. Updating strategies are adopted as follows:

$$\xi_j\left(A_{\eta_i}\right)(t+\tau) = (1-\rho)\,\xi_j\left(A_{\eta_i}\right)(t) + \triangle\xi_j\left(A_{\eta_i}\right) \tag{3}$$

where ρ $(0 \le \rho \le 1)$ is pheromone volatilization coefficient; $1-\rho$ is pheromone residual degrees.

$$\triangle\xi_j\left(A_{\eta_i}\right) = \sum_{k=1}^{Q}\triangle{\xi_j}^k\left(A_{\eta_i}\right) \tag{4}$$

where ${\xi_j}^k(A_{\eta_i})$ is the increment of the pheromone of ant k during the time interval τ.

$$\triangle{\xi_j}^k\left(A_{\eta_i}\right) = \begin{cases} \dfrac{C}{L_k} & \text{ant } k \text{ selects the } j \text{ element of the set } A_{\eta_i} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where C is a constant, and it represents the sum of the pheromones released by the ants after the completion of a cycle. L_k indicates the total path length of ant k in this foraging cycle.

Step 4 (repeating step 2 to step 3). When all the ants select the same path or reach the maximum number of iterations $N_{\max}$, the algorithm ends.

3.2. Improved Ant Colony Optimization Algorithm (IACO). In traditional ACO algorithm, the initial pheromone concentration of each location is the same value. These ants randomly choose the initial path to forage with equal probability strategy. Since the descendant ants choose the paths according to the cumulative pheromone concentration of the previous generation ant colony, it may appear that the number of ants on the non-global optimal paths is much more than that on the optimal path at the initial selection. This initial random selection strategy can increase the possibility of non-optimal paths selection. The algorithm is easy to fall into local optimum.

Here, we are focusing on revising the pheromone update strategy of traditional ant colony algorithm. In this paper, the traditional route selection process of ant colony is divided into three stages: pheromone pure increase stage, pheromone volatilization and accumulation stage, and pheromone doubling stage. The path selection strategies of three stages are shown as (6), (3), and (7).

The strategy of pheromone pure increase stage is formed as follows:

$$\xi_j\left(A_{\eta_i}\right)(t+\tau) = \xi_j\left(A_{\eta_i}\right)(t) + \triangle\xi_j\left(A_{\eta_i}\right) \tag{6}$$

where the formula for calculating $\triangle\xi_j(A_{\eta_j})$ is like (4).

The strategy of pheromone volatilization and accumulation stage is the same as the traditional ant colony routing strategy. The strategy is described in (3).

The strategy of pheromone doubling stage is expressed as follows:

$$\xi_j\left(A_{\eta_i}\right)(t+\tau) = (1-\rho)\,\xi_j\left(A_{\eta_i}\right)(t) + \gamma\triangle\xi_j\left(A_{\eta_i}\right) \tag{7}$$

where parameter γ is the doubling factor of process pheromone.

The procedure of IACO algorithm is as shown in Figure 1.

The pheromones are only accumulated and not volatilized in the previous N_1 iterations of algorithm. This stage is defined as the stage of pure pheromone addition. The total amount of pheromone increases $\triangle\xi_j(A_{\eta_j})$ during the time τ. Furthermore, this value is inversely proportional to the length of ants walking through the path. This means that the concentration of pheromones of shorter paths increases more during this process. The difference of cumulative pheromone concentration between the non-optimal paths and the optimal path is reduced. This strategy improves the possibility of optimal paths selection. Since the pheromone volatilization factor is eliminated in the pure increase stage of pheromone, the convergence speed of the algorithm slows down. In the following two stages, we adopt two strategies to accelerate the convergence of this algorithm. In the middle of the iterations, the traditional pheromone updating strategy is adopted. The strategy combines pheromone volatilization and accumulation. While speeding up the volatilization of pheromones, the possibility of non-optimal paths can be taken into account. In the last stage of the iteration, the process pheromone doubling factor is introduced when the number of iterations reaches N_2. The growth rate of pheromone is further improved, and the search speed of the optimal solution is also accelerated. This stage is called the pheromone doubling stage.

3.3. BP Neural Network. BP neural network [28] is a multilayer feedforward artificial neural network. This network simulates the cognitive process of human brain neurons to knowledge. BP algorithm updates the weights and thresholds of the network continuously by forward and backward information transmission and error back propagation. The network is trained several times by both positive and negative processes until the minimum error is satisfied. BP network has strong self-learning and fault tolerance properties. Figure 2 shows the topology structure of three layers BP network.

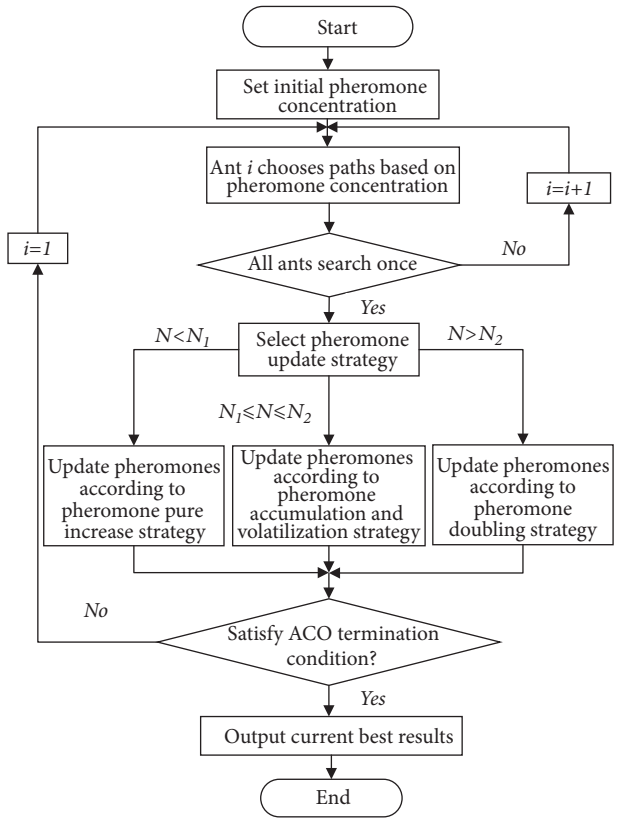


Figure 1: Flowchart of IACO algorithm.

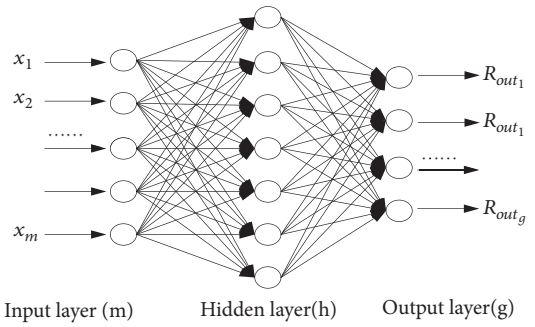


Figure 2: The topology structure of BP neural network.

BP network fitness function can be described as follows:

$$E\left(\eta_{ij}, \eta'_{qk}, \alpha_j, \beta_k\right) = \frac{1}{2}\sum_{p=1}^{P}\sum_{k=1}^{g}\left(R_k^p - R_{out_k}^p\right)$$
$$\eta_{ij} \in R^{m\times h},\ \eta'_{qk} \in R^{h\times l},\ \alpha_j \in R^{h\times 1},\ \beta_k \in R^{g\times 1} \tag{8}$$

where m is the number of nodes in the input layer, h is the number of hidden layer nodes, and g is the number of nodes in the output layer. P is the total number of training samples. η_{ij} and η'_{qk} are the network weight from node i in the input layer to node j in the hidden layer and the network weight from node q in the hidden layer to node k in the output layer, respectively. α_j and β_k are the threshold of node j in the hidden layer and the threshold of node k in the output layer, respectively. $R_{out_k}^p$ is the network output of training sample p. R_k^p is the expected value of sample p. BP network adjusts the error function $E(\eta_{ij}, \eta'_{qk}, \alpha_j, \beta_k)$ through bidirectional training to meet the error requirements, determines the network parameters, and obtains a stable network structure.

3.4. IACO-BP Algorithm. When BP neural network is used to predict, the initial weights and thresholds of each layer are given randomly. This method brings a great fluctuation in forecasting the same data. IACO-BP algorithm maps the optimal solution generated by IACO algorithm to the weights and thresholds of BP neural network. This strategy solves

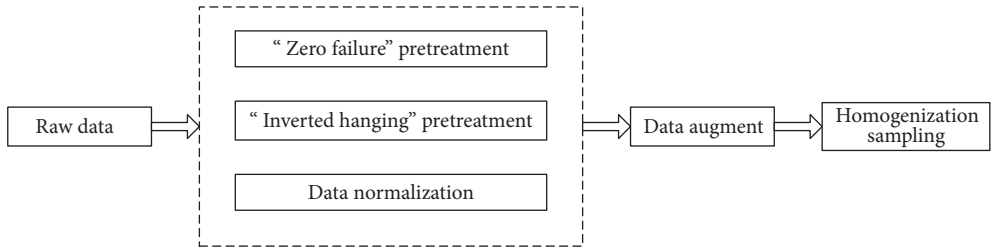


Figure 3: The process of data pretreatment and extraction.

largely the randomness of initial weights and thresholds of BP network. IACO-BP algorithm improves effectively the accuracy and stability of BP network prediction. The specific steps of the algorithm are demonstrated as follows.

Step 1 (normalizing the sample data).

$$\widetilde{x}_i = \frac{x_i - x_{\min}}{0.5\left(x_{\max} - x_{\min}\right)} - 1 \tag{9}$$

where $x_{\max}$ and $x_{\min}$ are the maximum value and the minimum value in the sample data, respectively. x_i is the ith input variable value. The range of $\widetilde{x}_i$ is in the interval $[-1, 1]$.

Step 2 (establishing three-layer network topology $m \times h \times g$). Determine the value of m, h, g.

Step 3 (initializing ant colony parameters). Set the parameters: M, Q, a, C, ρ, N_1, N_2, $N_{\max}$ and dimension $m \times h + h \times g + h + g$ of BP neural network parameters.

Step 4 (starting ants). Each ant starts from a location element in the set A_{η_1} to find the shortest path. The pheromone is updated according to the improved ant colony three-stage pheromone update strategy.

Step 5. Repeat Step 4, until all ants choose the same path or reach the maximum number of iterations, and turn to Step 6.

Step 6. Map the global optimal value in Step 5 to the initial weights and thresholds of BP neural network. Output the predicted value of the network. Calculate the fitness function $E(\eta_{ij}, \eta'_{qk}, \alpha_j, \beta_k)$. Propagate the error back to the output layer and adjust the weights and thresholds. Repeat the above process, until the terminating condition of BP algorithm ε_{BP} is satisfied or reaches BP maximum iteration $N'_{\max}$.

Step 7. Reverse the test data and restore the form of the test data.

$$y = 0.5\left(\overline{y}_i + 1\right)\left(y_{\max} - y_{\min}\right) + y_{\min} \tag{10}$$

where $y_{\max}$ and $y_{\min}$ are the maximum value and the minimum value of the network output data, respectively. $\overline{y}_i$ is the ith output value.

The framework of IACO-BP algorithm is as shown in Algorithm 1 .

4. Data Collection and Pretreatment

4.1. Data Set. In this paper, the statistical data of ammunition failure numbers are selected as sample data at different temperatures, humidity, and storage period under natural storage conditions. The data are shown in Table 1. We employ ammunition failure data to train ammunition failure prediction model. Ammunition reliability is indirectly predicted by ammunition failure numbers. Each sample capacity in data set is 10.

In Table 1, the unit of temperature is the international unit, Kelvin. It is represented by the symbol K. Relative humidity $RH\%$ record is used to indicate the percentage of saturated water vapor content in the air under the same condition.

4.2. Data Pretreatment. The experimental samples collected at the scene usually contain samples with "zero failure" and "inverted hanging" samples. These samples affect the accuracy of ammunition reliability prediction. It is necessary to pre-treat the data before the simulation experiments. Bayes estimation method is applied to treat "Zero failure" and "inverted hanging" data. Method of artificially filling noise signal is used to augment small sample size. Homogenization sampling is introduced to extract training samples and test samples. Figure 3 shows the process of data pretreatment and extraction.

(1) "Zero Failure" Data Pretreatment. In the random sampling of small sample ammunition life test, it may appear that all the samples are valid. This means that the ammunition failure number is zero. In reliability data prediction, since zero failure data provides less effective information, it needs to be pretreated. Aiming at the "zero failure" problem of samples No. 1, No. 9, and No. 29 in the data set, the increment function of reliability p is selected as the prior distribution of p, that is, $\pi(p) \propto p^2$. The posterior distribution of p is calculated based on Bayes method. Zero failure data are treated.

When zero failure occurs, the likelihood function of p is $L(0/p) = p^2$. By Bayes estimation method, the reliability $\widetilde{p}$ is obtained under the condition of square loss, as shown in (11). The zero failure f_i is transformed into $\widetilde{f}_i$, as described in (12).

$$\widetilde{p} = \int_0^1 ph\left(\frac{p}{n_i}\right)dp \tag{11}$$

$$\widetilde{f}_i = n_i\left(1 - \widetilde{p}\right) \tag{12}$$

Table 1: Raw data set.

Number	Temperature (K)	Humidity (RH%)	Period (Year)	Failure number
1	293	40%	3	0
2	293	40%	5	1
3	293	40%	8	1
4	293	40%	12	2
5	293	45%	5	1
6	293	45%	8	1
7	293	45%	15	2
8	293	45%	20	2
9	298	35%	3	0
10	298	35%	6	1
11	298	35%	12	2
12	298	35%	25	3
13	298	40%	5	1
14	298	40%	10	2
15	298	40%	12	1
16	298	40%	20	3
17	303	40%	7	1
18	303	40%	10	1
19	303	40%	12	1
20	303	40%	20	2
21	303	45%	5	1
22	303	45%	10	1
23	303	45%	15	2
24	303	45%	25	3
25	308	50%	8	1
26	308	50%	12	2
27	308	50%	15	1
28	308	50%	20	3
29	308	55%	2	0
30	308	55%	5	1
31	308	55%	15	2
32	308	55%	20	2

After "zero failure" processing, the failure numbers of samples No. 1, No. 9, and No. 29 in the data set are all converted to 0.714286.

(2) "Inverted Hanging" Data Pretreatment. Theoretically, under the same storage conditions, the number of ammunition failures with longer storage time is relatively large. During the same storage time, the number of ammunition failures with poor storage conditions is relatively large. The failure numbers of samples No.14 and No.15 and samples No.26 and No.27 are arranged inversely with the storage time, that is, the "inverted hanging" phenomenon. It needs to be "inverted hanging" treated.

Two sets of inverted data in the sample can be used to modify p_i to t_i time by selecting t_j time failure data p_j as the reference data in the corresponding period (t_j, t_i). Since both inverted samples are not at the end of the experiment at t_i time, the uniform distribution $U(p_j, p_i)$ is taken as a prior distribution of p_i. p_i is revised to $\widetilde{p}_i$.

$$\widetilde{p}_i = \frac{\int_{p_j}^{p_i+1} p^{f_{j+1}+1} (1-p)^{s_{j+1}} \, dp}{\int_{p_j}^{p_i+1} p^{f_{j+1}} (1-p)^{s_{j+1}} \, dp} \tag{13}$$

where s_j is the unfailing number in the sample. $\widetilde{p}_i$ is the failure probability corrected after t_i times. The corrected number of failures is $\widetilde{f}_i = n_i \widetilde{p}_i$. After "inverted hanging" pretreatment, the failure number of samples No.15 and No.27 in the data set is converted to 2. The "inverted hanging" phenomenon is eliminated. The data after zero failure and inverted hanging pretreatment are shown in Table 2.

(3) Normalization of Data. Temperature, humidity, and storage period in the data set are corresponding to different quantity rigidity and quantity scale. It is necessary to normalize

```
{ Initialize the ant colony parameters;
   begin:
      while(ants_i<ants_num)
         {ants_i chooses paths based on pheromone concentration;
          ants_i ++; }
            {  //Select different strategies in stages
                 if(iterations_num<N1)
                   {Update pheromones according to pheromone pure increase strategy }
               if(N1<iterations_num<N2)
                   {Update pheromones according to pheromone accumulation and
volatilization strategy }
               if(iterations_num>N2)
                   {Update pheromones according to pheromone doubling strategy }
              }
        if (current_error>margin_error)
              { ants_i=1;
goto begin;   }
        else
             { Input ant colony result is the initial solution of neural network;
               neural network training;
               resulte = neural network training results; }
        return resulte;}
```

Algorithm 1: IACO-BP (improved ant colony optimization Bp).

the data. The normalized data can eliminate the influence of variable dimension and improve the accuracy of evaluation. Normalization method is denoted as (9).

(4) Small Sample Data Augment. Since the particularity of ammunition samples, the problem of small sample size is usually encountered in the study of ammunition reliability prediction. In this paper, we add noise signals into the input signals after normalization. The method is applied to simulate the sample data and increase the sample size. The generalization ability of BP network is enhanced.

The number of samples is set to n_s, and the sample k is recorded as $p^k = (x^k, R^k_{out})$, $k = 1, 2, \cdots, n_s$. x^k is the input signal vector for the sample k, and R^k_{out} is the output signal of the sample k. The noise vector δ^k is added to the input signal. The output signal remains unchanged. The new sample becomes $\widetilde{p}^k = (x^k + \delta^k, R^k_{out})$. Set $\delta^k = 0.001$. Firstly, three components of input signal vector are added with noise δ^k, respectively. Then, two of the three components are grouped together, respectively. They are added with noise δ^k. Finally, these three components are added with noise δ^k, respectively. 32 data sets are extended to 256 data sets. The increase of sample size provides data support for BP neural network to mine the potential nonlinear laws in ammunition sample data. The smoothness of network training curve and the stability of network structure are improved.

4.3. Homogenization Sampling. In practical application scenarios, data have aggregation phenomenon. Aggregation phenomenon refers that a certain type of data set is distributed in a specific location of the data set. During analyzing data, it may occur that the sample data cannot represent the overall data. In this paper, a homogenization sampling data extraction method is proposed. Training samples and test samples are selected in the whole range. The bias of sample characteristics caused by aggregation sampling is removed. Ensure that the sample can maximize the overall data characteristics.

The implementation of homogenization sampling is as follows: Assuming that there are N sets of sample data, X sets of data are constructed as training samples, and Y sets are constructed as test data sets. Every N/Y-1 (if N/Y is a decimal, take a whole downward) data is test data until it is countlessly desirable. The remaining N-Y data are used as training samples. The homogenization sampling rule is shown in Figure 4.

It is known that the data pretreated in Section 4.2 are homogenized. In N=256 group samples, Y=10 group samples are taken as test samples. According to the homogenization sampling rule, one sample is taken from every $[N/Y]$-1=24 groups as the test sample, and the rest groups are as the training sample. Table 3 demonstrates four ammunition storage model precisions. These precisions are obtained based on homogenization sampling data and non-homogenization sampling data, respectively.

Table 3 shows the model accuracy of ammunition storage reliability prediction based on homogenization sampling much higher than that based on non-homogenization sampling. All algorithms in this paper are modeled and analyzed on the basis of homogenization sampling.

5. The Simulation Experiments

In this section, we show the specific parameter settings and the results of the simulation experiment. The mean square

Table 2: Data after zero failure and inverted hanging treatment.

Number	Temperature (k)	Humidity (RH)	Period (Year)	Failure number
1	293	40%	3	0.714286
2	293	40%	5	1
3	293	40%	8	1
4	293	40%	12	2
5	293	45%	5	1
6	293	45%	8	1
7	293	45%	15	2
8	293	45%	20	2
9	298	35%	3	0.714286
10	298	35%	6	1
11	298	35%	12	2
12	298	35%	25	3
13	298	40%	5	1
14	298	40%	10	2
15	298	40%	12	2
16	298	40%	20	3
17	303	40%	7	1
18	303	40%	10	1
19	303	40%	12	1
20	303	40%	20	2
21	303	45%	5	1
22	303	45%	10	1
23	303	45%	15	2
24	303	45%	25	3
25	308	50%	8	1
26	308	50%	12	2
27	308	50%	15	2
28	308	50%	20	3
29	308	55%	2	0.714286
30	308	55%	5	1
31	308	55%	15	2
32	308	55%	20	2

Table 3: Precision comparison.

Method	IACO-BP	ACO-BP	PSO-BP	BP
Non-homogenization sampling	0.8075	0.8201	0.3259	1.2656
Homogenization sampling	6.88e-04	4.29e-03	0.0377	0.0811

error (MSE) and mean absolute percentage error (MAPE) of performance indicators are used to evaluate the prediction effect of the model. The IACO-BP model shows its implement performance in comparison with BP, PSO-BP, and ACO-BP models.

$$MSE = \frac{1}{N}\sum_{k=1}^{N}\left|R_k - \widehat{R}_{out_k}\right| \tag{14}$$

$$MAPE = \frac{1}{N}\sum_{k=1}^{N}\left|\frac{R_k - \widehat{R}_{out_k}}{R_k}\right| \times 100\% \tag{15}$$

where R_k and $\widehat{R}_{out_k}$ are expected reliability and actual reliability. N is the total number of data used for reliability prediction and comparison.

5.1. Parameter Settings. In this paper, the storage temperature, humidity, and storage period of ammunition under natural storage are used as input variables of neural network to predict the number of ammunition failure. The storage reliability of ammunition is indirectly predicted by the number of ammunition failure. The number of input nodes in BP network is m=3, and the number of network output nodes is g=1. According to empirical formula $h = \sqrt{m+g} + c$, the

Table 4: The results of trial and error.

Hidden layer numbers	2	3	4	5	6	7	8	9	10	11	12	13
The mean of MSE	0.226	0.165	0.198	0.269	0.357	0.266	0.413	0.205	0.288	0.026	0.378	0.566

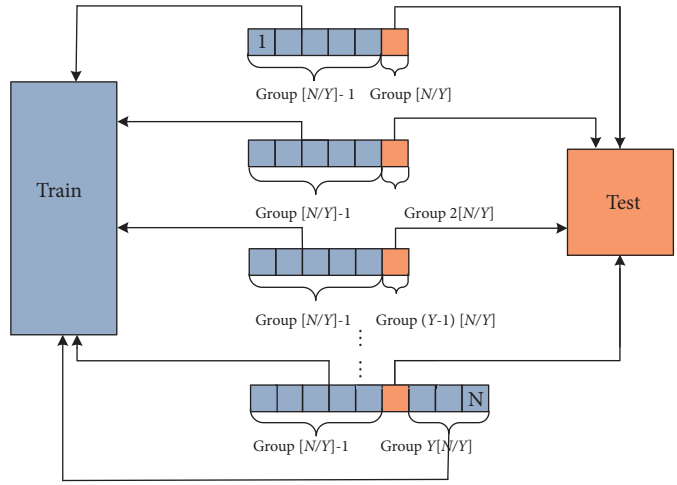


Figure 4: Schematic diagram of homogenization sampling.

value range of hidden layer nodes is [2, 12]. c is the constant in the interval [1, 10]. In order to avoid the contingency of the result caused by the instability of the neural network, 10 experiments are conducted on the same number of hidden layer nodes. The value of MSE of the predicted and true values is observed. The experimental results of trial and error are shown in Table 4.

Table 4 indicates that the mean value of MSE is the smallest when the number of hidden layer nodes is 11. Therefore, the neural network topology in this paper is $3 \times 11 \times 1$.

The activation function of the hidden layer of BP network is "tansig", the activation function of the output layer is "purelin", and the training function of the network is "trainlm". The learning rate is set to 0.1. The expected error is $\varepsilon_{BP} = 0.001$. The weights and thresholds parameters dimension of BP neural network are 56. Each parameter is randomly assigned to 20 values in the internal $[-1, 1]$. The number of ants is $Q = 80$. Pheromone initial value is set to $a = 1$. The coefficient of pheromone volatilization is $\rho = 0.1$. The increment of pheromone is $C = 1$. The number of terminations of the pheromone pure increase stage is set to 200, the number of initial iterations of the pheromone doubling stage is set to 500, and the number of iterations between 200 and 500 is the stage of pheromone volatilization and accumulation. The doubling factor of process pheromone is $\gamma = 2$. Maximum iteration of ant colony is set to $N_{\max} = 800$.

The algorithm is coded by MATLAB 2016a. The failure numbers of ammunition storage are generated by BP neural network algorithm, ACO-BP algorithm, PSO-BP algorithm, and IACO-BP algorithm, respectively. Then the results are compared and analyzed.

5.2. Results Analysis. In order to evaluate performance of the four algorithms, 20 simulation tests are conducted on each algorithm to obtain the value of the prediction error evaluation index MSE and MAPE. The arithmetic mean values of MSE and MAPE in 20 tests are used to characterize the accuracy of four algorithms. The variance, standard deviation, and range of MSE and MAPE are applied to evaluate the stability of algorithms. The results are shown in Tables 5 and 6.

The mean value of MSE of IACO-BP algorithm is 1.2e-03, which is 92%, 95%, and 99% lower than that of ACO-BP, PSO-BP, and BP algorithm, respectively. The mean value of MAPE of IACO-BP algorithm is 2.1e+00, which is 66%, 79%, and 83% lower than that of ACO-BP, PSO-BP, and BP algorithm, respectively. It can be seen that, in the four intelligent algorithms, the IACO-BP algorithm has the highest accuracy. For IACO-BP algorithm, the variance of MSE is 4.6e-07, the mean standard deviation of MSE is 6.8e-04, and the range of MSE is 2.2e-03. Each stability index of IACO-BP is the minimum value of corresponding index in the four algorithms. Hence, IACO-BP algorithm has better stability. The MSE value of the BP algorithm fluctuates significantly higher than the other three optimized BP algorithms, as shown in Figure 5. IACO-BP algorithm is the best stability among these three intelligent BP optimization algorithms, as shown in Figure 6. Table 6 shows the statistical value of the number of iterations obtained from 10 random trainings of four networks. The maximum number of iterations in the networks is 10000, and the network accuracy is 0.001. As shown in Table 6, the 10 simulations of PSO-BP network fail to satisfy the accuracy requirement of the network during 10000 iterations and terminate the algorithm in the way of achieving the maximum number of iterations. The average number of iterations in IACO-BP network is 339.8, which is the smallest among these four algorithms. The MSE of

Table 5: Simulation results.

Serial ID	IACO-BP		ACO-BP		PSO-BP		BP	
	MSE	MAPE	MSE	MAPE	MSE	MAPE	MSE	MAPE
1	2.3e-03	3.0e+00	4.8e-04	1.6e+00	3.8e-02	1.3e+01	3.8e-02	9.8e+00
2	4.3e-04	1.4e+00	3.3e-03	3.7e+00	2.6e-02	1.2e+01	3.8e-02	1.0e+01
3	9.4e-04	2.1e+00	3.2e-03	4.1e+00	7.9e-03	5.4e+00	1.0e-02	6.1e+00
4	1.1e-03	2.2e+00	3.4e-03	4.3e+00	2.1e-02	8.8e+00	1.0e-02	6.9e+00
5	2.1e-03	2.7e+00	9.1e-03	6.2e+00	1.2e-02	9.0e+00	7.7e-03	5.1e+00
6	1.1e-03	2.3e+00	2.6e-02	8.3e+00	4.0e-02	1.3e+01	1.2e-02	7.8e+00
7	1.3e-03	2.0e+00	4.6e-02	1.3e+01	1.1e-02	7.0e+00	1.1e+00	4.1e+01
8	2.0e-03	2.7e+00	4.0e-02	1.1e+01	1.7e-02	7.8e+00	1.8e-03	2.6e+00
9	6.7e-04	2.0e+00	2.1e-03	2.6e+00	3.7e-02	1.1e+01	2.7e-03	3.8e+00
10	2.5e-03	2.4e+00	2.0e-03	2.6e+00	1.1e-02	8.7e+00	6.1e-02	1.4e+01
11	1.0e-03	1.8e+00	4.3e-02	1.2e+01	4.3e-02	1.3e+01	1.1e+00	4.1e+01
12	4.3e-04	1.6e+00	1.4e-03	2.5e+00	9.0e-03	7.1e+00	5.4e-03	4.6e+00
13	3.4e-04	9.7e-01	3.7e-03	3.6e+00	2.2e-02	9.0e+00	2.6e-02	1.1e+01
14	2.0e-03	2.7e+00	1.3e-03	2.1e+00	2.1e-02	8.5e+00	4.1e-02	1.1e+01
15	1.7e-03	3.0e+00	4.5e-04	1.4e+00	7.8e-02	1.9e+01	1.6e-03	2.4e+00
16	1.1e-03	2.1e+00	2.1e-03	2.7e+00	7.6e-02	1.7e+01	4.4e-03	4.3e+00
17	4.9e-04	1.6e+00	4.0e-03	4.4e+00	2.6e-02	9.2e+00	8.9e-03	6.1e+00
18	8.7e-04	1.8e+00	1.2e-02	6.4e+00	3.4e-03	3.7e+00	1.1e+00	4.1e+01
19	4.9e-04	1.5e+00	4.6e-02	1.2e+01	8.8e-03	6.4e+00	8.3e-03	4.4e+00
20	1.3e-03	2.2e+00	2.0e-02	8.4e+00	2.3e-02	9.4e+00	3.5e-02	1.2e+01
mean	1.2e-03	2.1e+00	1.5e-02	6.1e+00	2.7e-02	9.9e+00	1.7e-01	1.2e+01
Variance	4.6e-07	3.1e-01	2.9e-04	1.4e+01	4.3e-04	1.5e+01	1.5e-01	1.7e+02
Standard deviation	6.8e-04	5.6e-01	1.7e-02	3.7e+00	2.1e-02	3.8e+00	3.8e-01	1.3e+01
Range	2.2e-03	2.0e+00	4.6e-02	1.1e+01	7.5e-02	1.6e+01	1.1e+00	3.9e+01

Table 6: Comparison of iterations.

Number of iterations	IACO-BP	ACO-BP	PSO-BP	BP
Mean	339.8	438.7	10000	708.3
MSE	144.88	901.17	0	1110.08

Table 7: Comparison of reliability prediction.

Actual reliability	0.8	0.9	0.7	0.9	0.9	0.9286
IACO-BP	0.8012	0.8956	0.6984	0.9041	0.9059	0.9256
ACO-BP	0.7967	0.9112	0.7020	0.8949	0.9000	0.9293
PSO-BP	0.8128	0.8924	0.7018	0.8656	0.8926	0.9341
BP	0.8025	0.8985	0.7019	0.8470	0.9012	0.9286

iteration is 144.88, which is also the smallest value among these three algorithms which satisfy the iterative precision termination procedure.

Six failure data samples are randomly selected from the field records, and the failure numbers are predicted using by the four trained networks. The storage reliability of ammunition is indirectly predicted by using (1). The results are described in Table 7 and Figure 7. The indirect prediction of ammunition storage reliability model based on IACO-BP algorithm has the best fitting degree and the highest accuracy among the four algorithms.

6. Conclusion

In this paper, a prediction model of the ammunition storage reliability is considered under natural storage conditions. IACO-BP neural network algorithm is proposed to solve this problem. Reliability is indirectly predicted by the number

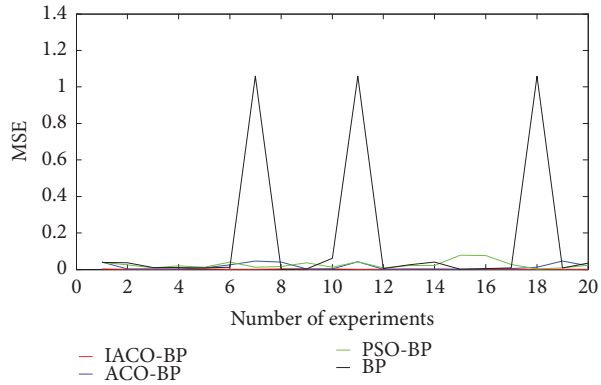


Figure 5: MSE curves of four algorithms.

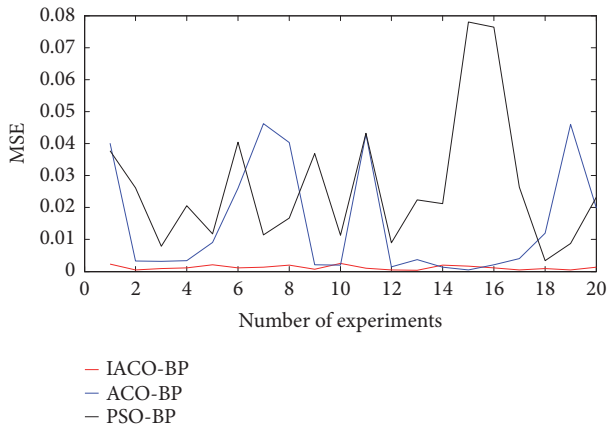


Figure 6: MSE curves of three optimizing BP algorithms.

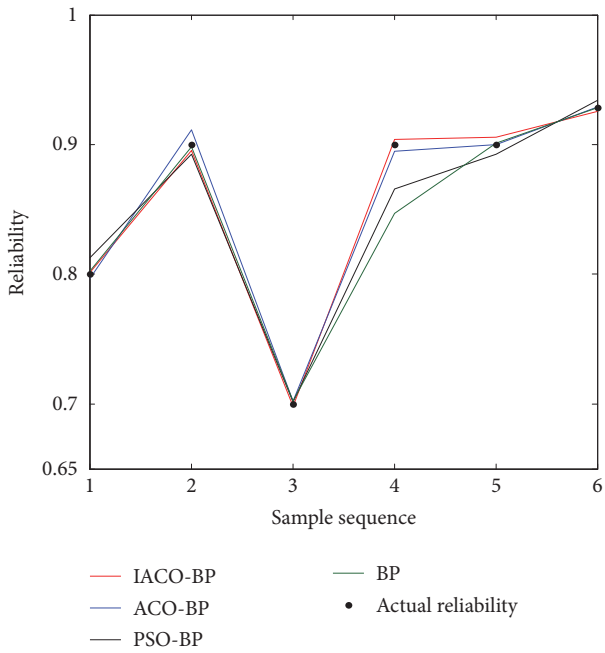


Figure 7: Reliability prediction curves of four algorithms.

of ammunition failures obtained by IACO-BP ammunition failure number prediction model. In order to improve the accuracy and stability of network prediction, the data pretreatment and algorithm are promoted. In data pretreatment aspect, the standardization and rationality of the data sample are derived by using the methods of "zero failure" pretreatment, "inverted hanging" pretreatment, small sample data augmentation, and homogenization sampling. In algorithm aspect, we improve a pheromone updating strategy from the traditional ant colony to a three-stage updating strategy of pheromone pure increment, volatilization and accumulation, and doubling. Compared with BP, PSO-BP, and ACO-BP model, the experimental results prove that IACO-BP model has great advantages in precision, stability, and iteration times.

Data Availability

The raw data used to support the findings of study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by Open Foundation of Science and Technology on Electro-Optical Information Security Control Laboratory (Grant no. 61421070104), Natural Science Foundation of Liaoning Province (Grant no. 20170540790), and Science and Technology Project of Educational Department of Liaoning Province (Grant no. LG201715).

References

[1] Z. Xiaonan, Y. Junfeng, D. Siliang, and H. Shudong, "A New Method on Software Reliability Prediction," *Mathematical Problems in Engineering*, vol. 2013, Article ID 385372, 8 pages, 2013.

[2] P. Gao, L. Xie, and W. Hu, "Reliability and random lifetime models of planetary gear systems," *Shock and Vibration*, vol. 2018, Article ID 9106404, 12 pages, 2018.

[3] S. Wang, "Reliability model of mechanical components with dependent failure modes," *Mathematical Problems in Engineering*, vol. 2013, Article ID 828407, 6 pages, 2013.

[4] X. Chen, Q. Chen, X. Bian, and J. Fan, "Reliability evaluation of bridges based on nonprobabilistic response surface limit method," *Mathematical Problems in Engineering*, vol. 2017, Article ID 1964165, 10 pages, 2017.

[5] G. Peter O'Hara, "Dynamic analysis of a 155 mm cannon breech," *Shock and Vibration*, vol. 8, no. 3-4, pp. 215–221, 2001.

[6] Z. Z. Muhammad, I. B. Shahid, I. Tauqeer, Z. E. Syed, and F. P. Zhang, "Nonlinear material behavior analysis under high compression pressure in dynamic conditions," *International Journal of Aerospace Engineering*, vol. 2017, Article ID 3616932, 15 pages, 2017.

[7] B. Zheng, H. G. Xu, and Z. B. Jiang, "An estimation method of ammunition storage life based on poisson process," *Acta Armamentarli*, vol. 26, no. 4, pp. 528–530, 2005.

[8] L. J. Gullo, A. T. Mense, J. L. Thomas, and P. E. Shedlock, "Models and methods for determining storage reliability," in *Proceedings of the 2013 Annual Reliability and Maintainability Symposium (RAMS)*, pp. 1–6, Orlando, Fla, USA, 2013.

[9] C. Su, Y.-J. Zhang, and B.-X. Cao, "Forecast model for real time reliability of storage system based on periodic inspection and maintenance data," *Eksploatacja I Niezawodnosc-Maintenance and Reliability*, vol. 14, no. 4, pp. 342–348, 2012.

[10] Z. Liu, X. Ma, and Y. Zhao, "Storage reliability assessment for missile component with degradation failure mode in a temperature varying environment," *Acta Aeronautica Et Astronautica Sinica*, vol. 33, no. 9, pp. 1671–1678, 2012.

[11] X. S. Si, C. H. Hu, X. Y. Kong, and D. H. Zhou, " A residual storage life prediction approach for systems with operation state switches," *IEEE Transactions on Industrial Electronics*, vol. 61, no. 11, pp. 6304–6315, 2014.

[12] Z. Wang, C. Hu, W. Wang, Z. Zhou, and X. Si, "A case study of remaining storage life prediction using stochastic filtering with the influence of condition monitoring," *Reliability Engineering & System Safety*, vol. 132, pp. 186–195, 2014.

[13] Y. Zhang, M. Zhao, S. Zhang, J. Wang, and Y. Zhang, "An integrated approach to estimate storage reliability with initial failures based on E-Bayesian estimates," *Reliability Engineering & System Safety*, vol. 159, pp. 24–36, 2017.

[14] B. Zheng and G.-P. Ge, "Estimation of fuze storage life based on stepped stress accelerated life testing," *Transactions of Beijing Institute of Technology*, vol. 23, no. 5, pp. 545–547, 2003.

[15] W. B. Nelson, "A bibliography of accelerated test plans," *IEEE Transactions on Reliability*, vol. 54, no. 2, pp. 194–197, 2005.

[16] J. L. Cook, "Applications of service life prediction for US army ammunition," *Safety and Reliability*, vol. 30, no. 3, pp. 58–74, 2010.

[17] Z. G. Shen, J. C. Yuan, J. Y. Dong, and L. Zhu, "Research on acceleration factor estimation method of accelerated life test of missile-borne equipment," *Engineering & Electronics*, vol. 37, no. 8, pp. 1948–1952, 2015.

[18] Z.-C. Zhao, B.-W. Song, and X.-Z. Zhao, "Product reliability assessment by combining accelerated degradation test and accelerated life test," *System Engineering Theory and Practice*, vol. 34, no. 7, pp. 1916–1920, 2014.

[19] X. Y. Zou and R. H. Yao, "Small sample statistical theory and IC reliability assessment," *Control and Decision*, vol. 23, no. 3, pp. 241–245, 2008.

[20] M. Zhang and W. Chen, "Hot spot data prediction model based on wavelet neural network," *Mathematical Problems in Engineering*, vol. 2018, Article ID 3719564, 10 pages, 2018.

[21] H. J. Chen, K. N. Teng, B. Li, and J. Y. Gu, "Application of neural network on missile storage reliability forecasting," *Journal of Projectiles, Rockets, Missiles and Guidance*, vol. 30, no. 6, pp. 78–81, 2010.

[22] J. Liu, D. Ling, and S. Wang, "Ammunition storage reliability forecasting based on radial basis function neural network," in *Proceedings of the 2012 International Conference on Quality, Reliability, Risk, Maintenance, and Safety Engineering*, pp. 599–602, Chengdu, China, 2012.

[23] Z. G. Tian, L. N. Wong, and N. M. Safaei, "A neural network approach for remaining useful life prediction utilizing both failure and suspension histories," *Mechanical Systems and Signal Processing*, vol. 24, no. 5, pp. 1542–1555, 2010.

[24] E.-Z. Zhang and Q.-W. Chen, "Multi-objective particle swarm optimization for uncertain reliability optimization problems," *Control and Decision*, vol. 30, no. 9, pp. 1701–1705, 2015.

[25] H. Gong, E. M. Zhang, and J. Yao, "BP neural network optimized by PSO algorithm on ammunition storage reliability prediction," in *Proceedings of the 2017 Chinese Automation Congress (CAC)*, pp. 692–696, Jinan, China, 2017.

[26] X. Liu, L. Fan, L. Wang, and S. Meng, "Multiobjective reliable cloud storage with its particle swarm optimization algorithm," *Mathematical Problems in Engineering*, vol. 2016, Article ID 9529526, 14 pages, 2016.

[27] Lin Yuan, Chang-An Yuan, and De-Shuang Huang, "FAACOSE: a fast adaptive ant colony optimization algorithm for detecting SNP epistasis," *Complexity*, vol. 2017, Article ID 5024867, 10 pages, 2017.

[28] F. P. Wang, "Research on application of big data in internet financial credit investigation based on improved GA-BP neural network," *Complexity*, vol. 2018, Article ID 7616537, 16 pages, 2018.

Hindawi
Complexity
Volume 2019, Article ID 4164853, 17 pages
https://doi.org/10.1155/2019/4164853

Research Article

Green Start-Ups' Attitudes towards Nature When Complying with the Corporate Law

Rafael Robina-Ramírez,[1] **Antonio Fernández-Portillo**,[2] **and Juan Carlos Díaz-Casero**[1]

[1]*Business and Sociology Department, University of Extremadura, Avda de la Universidad s/n, 10071 Cáceres (Extremadura), Spain*
[2]*Finance and Accounting Department, University of Extremadura, Avda de la Universidad s/n, 10071 Cáceres (Extremadura), Spain*

Correspondence should be addressed to Rafael Robina-Ramírez; rrobina@unex.es

Received 1 December 2018; Revised 25 January 2019; Accepted 10 February 2019; Published 21 February 2019

Academic Editor: Rongqing Zhang

This paper examines how Spanish green start-ups develop improved attitudes towards nature. Despite the prevalence of theories based on green entrepreneurs, very little research has been conducted on how green attributes influence nature in the framework of the new Spanish Criminal Code in what concerns corporate compliance. One hundred and fifty-two start-ups were interviewed in 2018. Smart PLS Path Modelling has been used to build an interaction model among variables. The results obtained have a significant theoretical and practical implication since they add new findings to the current literature on how start-ups incorporate the recent Criminal Code in their environmental decisions. The model reveals a strong predictive power (R-squared = 77.9%). There are grounds to say that start-ups should implement effective monitoring systems and new organisational standards, in addition to developing surveillance measures to avoid unexpected sanctions.

1. Introduction

The paper delves into green entrepreneurs' attitudes in start-ups when complying with the environmental law in the context of the recent Spanish Criminal Code in what concerns corporate compliance. Regulative compliance explains the aim that organisations want to reach in their efforts to comply with environmental regulations.

It is estimated that 99.9% of Spanish organisations are SMEs, which means that there are 2,498,870 SMEs in Spain capable of delivering more sustainable goods and services through organic food production, fair trade, natural and handmade craft, sustainable tourism services, environmental consulting, green energy firms, etc. [1].

Spain has implemented several policies to enhance entrepreneurship in the past decades. Businesses are required to develop rapidly, reduce unemployment rates, and improve the deteriorated economy without harming the environment [1].

Since Brundtland wrote Our Common Future (1987) [2], companies have been aided to delve into the needs of the current time. It is proposed to look after the future generations in order to meet their own demands in their development efforts. The sustainable argument in reduced companies has gradually expanded [3] and is supported by empirical studies [4].

For decades, the emerging literature on green entrepreneurship in small companies has focused on environmentally oriented entrepreneurship achieving sustainable development [5]. The evolution of green entrepreneurs has recently translated into the appearance of many green start-ups in a wide variety of sectors, business strategies, and marketing targets. These entrepreneurs have created products and services to meet Green Opportunities and benefit nature at the same time [6]. Ivanko [7] added the social dimension linked to the environment to ascertain the problems suffered by the community. Social entrepreneurship brings a range of new social values to profits, environment, and fair trade [8].

Even though there is no agreement on the definition of green entrepreneurs [9], they are viewed by scholars as companies with the predisposition to pursue potential opportunities that produce both economic and ecological benefits through green activities.

Their green market orientation has defined their dual identity [10]. Thus, profit and environment may compete on equal terms. As for the way green companies operate in the market, their environmental orientation reflects green innovativeness, market proactiveness, and risk-taking [11, 12].

Considering these features, Walley and Taylor [13] highlighted three pillars of green entrepreneurs as three key features of sustainable development: economic prosperity based on the balance between economy and environment, the pursuit of environmental quality to protect nature through innovative and competitive processes, and the social and ethical consideration of the business culture focused on promoting Green Opportunities among companies.

Based on the literature review, four attributes were extracted from these pillars to define a model for creating environmental value for green start-ups [14, 15]. These attributes are Balanced Green Investment [16], environmental impact of green entrepreneurs [4, 17, 18], innovation and competitiveness [19–21], and Green Opportunities [22].

The attributes had to be assessed since a new regulatory compliance emerged in 2015 in Spain [23]. The new legal perspective relies on regulation and public concern to overcome substantial damage to the quality of air, soil or water, animals, or plants. In the context of the regulatory framework, effective monitoring systems not only improve the organisational management but also avoid unexpected sanctions [24].

The contribution of this paper is threefold: (1) it tries to ascertain whether the Environmental Value Model is empirically reflected in these attributes and validate the different opportunities for green start-ups; (2) it analyses whether the recent Spanish regulatory compliance influences green start-ups not only by imposing coercive rules but also by allowing them to design organisational norms of surveillance [25]; and (3) it explores whether the new regulatory compliance scenario affects green entrepreneurs' attitudes towards nature.

The relation between green entrepreneurs' attitudes and the recent pressure exerted by the Spanish regulatory compliance appears not to have been considered by green entrepreneurs. As a result, the paper examines what role green start-ups play in the new legal framework when complying with the Corporate Law to protect nature.

One hundred and fifty-two environmental green start-ups were involved in the study. The data collected was processed using SEM-PLS Path Modelling. Theoretical and practical implications were provided not only for environmental authorities but also for green entrepreneurs when avoiding unlawful environmental activity that damages nature. The research conducted further adds content to the current literature, particularly, about the findings of green start-ups when incorporating related measures into their organisational decisions.

The paper is structured as follows. First, green entrepreneurship is studied in the context of Spanish regulatory compliance, and then an Environmental Value Model for green entrepreneurs is proposed. In third place, the methodology used for this paper is explained. Fourth, results are drawn from the data collected. Finally, the discussion, conclusions, and limitations are presented.

2. Literature Review

2.1. Green Entrepreneurship in the Context of Spanish Regulatory Compliance

H_1: Environmental Corporate Compliance (ECC) Positively Influences Green Attitudes towards Nature (GAN) among Entrepreneurs. Since the industrial degradation occurred, environmental entrepreneurs emerged quite rapidly as new companies in the environmental sector and, more recently, as green start-ups [26]. Sustainable research in the field of entrepreneurship becomes crucial to avoid threats to nature [27]. In this framework, entrepreneurs have constantly desired to minimise their impact on the environment [28] seeking solutions to align economy with ecology [16]. The proposal is structured in ongoing process of improved actions and attitudes connected to the emerging environmental opportunities through organisational discoveries [9].

However, since the Spanish regulatory compliance emerged in 2015, a new perspective to deter environmental damage has been developed. The first environmental concern for nature came with the Spanish Constitution [29] and, later, the Spanish Criminal Code of 1983.

Administrative sanctions were established for individuals who committed unlawful environmental activities, such as causing emissions, discharges, extractions or excavations, vibration, injections, or deposits, in the atmosphere, soil, or terrestrial and marine waters (Article 173.1) [29].

Fortunately, the rules and regulations have gradually increased their protection over environment. The purpose has been to make an inhabitable environment place to live as well as to raise green awareness of sustainable principles among decision-makers [30].

To date, most ethical endeavour to defend the environment has not been consistent enough to adequately address environmental threats [31].

With the reform of the Spanish Corporate Governance Code [23] judges were allowed to hear cases not only about individuals having committed unlawful acts, but also about companies. The legal reform was based on transferring to companies the same rights and obligations as individuals. Although individuals are typically liable for committing environmental crimes towards nature, liability can now be transmitted to the enterprise (vicarious liability) [32].

On the one hand, the new legal framework is more severe than the previous one due to the new obligation for companies to comply with the Criminal Code and on the other hand, the legal reform allows companies to avoid sanctions by developing surveillance and control measures [23]. Such organisational standards must prevent information bias and foster green attitudes towards nature [33].

The strategy used to comply with this law should be designed through innovation and the creation of a specific department to deliver organisational measures [34, 35].

The process, however, requires developing surveillance protocols through innovation and creation procedures [36]

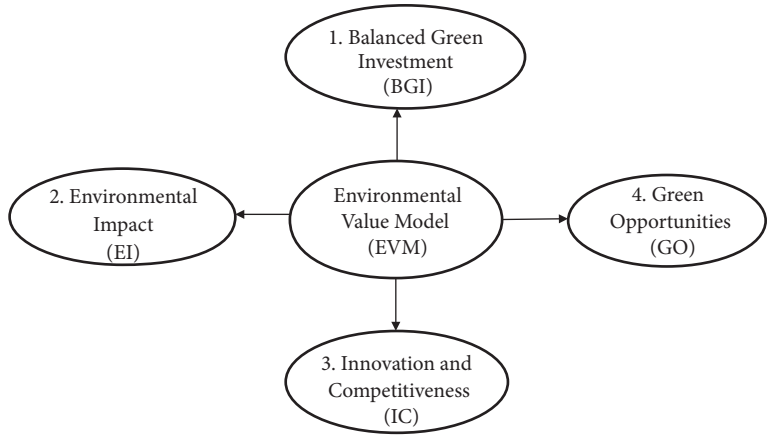


Figure 1: Environmental Value Model (EVM).

to balance companies' economic and ecologic activities [16]. Protocols alert green entrepreneurs about the risks that companies could be facing when their workers behave irresponsibly towards nature [37], which entails the company being seen as carrying out nonenvironmental and illegal activities [38].

H_2: Environmental Corporate Compliance (ECC) Positively Influences the Environmental Value Model (EVM). This new culture promoted by law has led to a new behavioural model that adds value to companies. The EVM prevents companies not only from committing unlawful environmental activity [39] but also from performing incorrect environmental procedures and the underlying uncertainty of reducing the risk of negative implications for nature [4].

Likewise, the preventive activities of the EVM for regulatory compliance require frequent assessment by establishing standards of procedures to ensure being up to date in the event of irresponsible behaviours leading to environmental crimes [40]. Moreover, it is also essential to settle disciplinary procedures through sanctions and prevent illegal conduct along the creation and innovation process [41].

Organisational standards of procedures are key factors of the environmental model. They have to be established by the compliance officer [42]. Such rules not only have to be supervised periodically [43] but also have to be communicated to the employees, e.g., by training programmes or spreading publications to explain these items and detect possible unlawful conduct empirically.

The following organisational decisions must be made in companies in their endeavour to comply with legal requirements, according to the Spanish Corporate Governance Code [23].

(i) A department must be created to supervise the compliance of green innovations, procedures, and services [34]. This department should establish a management model to avert unlawful environmental actions and allocate a corporate budget for surveillance practices within the company [44].

(ii) Training programmes on the protocols or procedures must be taught to appropriately implement surveillance measures in companies [36]. Protocols must be designed to report any prior unlawful behaviour within the company [38] and reveal any necessary changes to maintain legality in every business activity [40].

(iii) A disciplinary system must be implemented to address unlawful environmental actions and also the financial, managerial, and social obligations with the company and society [41].

For the three steps to be put into effect, recent corporate laws have enhanced several transparency and cooperation models within companies that have added value to environmental companies.

2.2. Environmental Value Model for Green Entrepreneurs (EVM)

H_3: The Environmental Value Model (EVM) Positively Influences Green Attitudes towards Nature (GAN). Innovations and business models have added increasing value to companies in the past decades. Emerging sustainable approaches such as fair trade, circular economy, or lowsumerism, have set up new trends [45] in response to demanding green consumers [46].

These models require defining key attributes to ascertain practical consequences in green entrepreneurs. Four attributes have been extracted from the literature review in this respect: (1) Balanced Green Investment, (2) environmental impact of green entrepreneurs, (3) innovation and competitiveness, and (4) Green Opportunities.

The contribution of this model is twofold: first, to find a predictive green model to delve into the empirical implications for Spanish green start-ups and, second, to analyse whether the balance between economy and sustainability is empirically relevant for Spanish green start-ups. Figure 1 shows the attributes of the model.

H_4: The Environmental Value Model (EVM) Positively Influences Balanced Green Investment (BGI). According to the OECD [47] sustainability creates conditions in which individuals and environment interact harmoniously. Sustainability allows reaching environmental, economic, and social standards to respect future generations.

In the context of environmental deterioration, sustainable management has become a pivotal issue [48]. Even though literature about damaging environmental outcomes is vast, there is a widespread ignorance towards the positive initiatives taken by green entrepreneurs to address such damage. Such ignorance is due to failures in coordinating expectations and preferences throughout the business models to reduce the actual high cost for companies [49]. Green models have introduced potential opportunities to produce both economic and ecological benefits through eco-friendly products and services [50]. How these green models influence environmental and financial performance remains unclear. While some studies have found a negative relationship between tangible-external green strategies and entrepreneurial performance [51], the impact of green entrepreneurship performance has also been positive in some cases [16].

The balance between investment for profit and to protect nature largely depends on the responsible performance of green entrepreneurs [52] and the voluntary disclosure of environmental standards [53]. Developing managerial standards to have a positive impact on the environment is precisely one of the core concepts of the Spanish regulatory compliance.

H_5: The Environmental Value Model (EVM) Positively Influences Innovation and Competitiveness (IC). The Environmental Value Model (EVM) is also associated with three features of entrepreneurs: innovativeness, proactiveness, and risk-taking [19].

First, innovativeness describes a tendency to deliver new ideas, engage in experimentation, and support creative processes. Innovation enables entrepreneurs to combine resources to launch new products or processes [20], to gain advantages over competitors [54], and to differentiate themselves from others [21]. Competitive advantages can also enhance companies' absorptive capacity by developing their ability to imitate advanced green technologies [55].

Second, proactiveness refers to the capacity to respond to customer needs by introducing green products, services, or technology [21]. Companies are facing growing pressure from customers with raising awareness of environmental issues. Proactive companies are likely to respond more quickly to the needs of customers than their competitors. Under the trend of customers' attitude towards green marketing, companies can reap the financial benefits of becoming a pioneer in green innovation practices.

Third, risk-taking is one of the primary personal attributes of entrepreneurs and it reflects the tendency to adopt an active stance when investing in projects [56]. Although the propensity of risk may bring fresh revenue, it is often associated with complex situations and uncertainties that can entail companies becoming trapped in changing circumstances [57].

H_6: The Environmental Value Model (EVM) Positively Influences Environmental Impact (EI). Green entrepreneurial activity can simultaneously foster economic and ecological benefits for society, by creating market opportunities and preventing environmental degradation [58] in two different ways. First, the entrepreneurial action may reduce environmental degradation and capture economic value by alleviating market failure [4] and using green technologies to decrease the consumption of water, electricity, coal, or oil [17]. Second, green entrepreneurs can reduce unhealthy damage to employees at work by decreasing the consumption of toxic substances [18] and protecting nature by applying corporate tax incentives to green companies [59].

H_7: The Environmental Value Model (EVM) Positively Influences Green Opportunities (GO). One way of building opportunities to expand environmental values within the company is by developing employee's green skills [22].

According to TECCe's argument, Green Opportunities facilitate the generation of new product processes [21]. Assessing potential opportunities and adopting eco-friendly technologies could very much help to overcome environmental market failures [60]. The electrical utility industry, for instance, has the possibility of taking more advantage out of wind power [10] and to make more green use of natural resources [61].

After describing the four green attributes that form the Environmental Value Model, the paper turns to the analysis of the methodology, results, discussion, and conclusions.

3. Research Methodology

3.1. Data Collection and Sample. The lack of an official list of Spanish environmental start-ups has hampered our endeavour to list and collect the data of SMEs. The research team chose green start-ups for two reasons: (1) The number of green start-ups has recently increased in Spain, and they are becoming a new phenomenon of environmental awareness and (2) given the lack of studies in the context of the recent Spanish regulatory compliance, this paper helps to shed light on the role green start-ups play and how the new regulation affects them.

242 online environmental green start-ups were initially identified in Spain. Following the literature review, the focus has been on those who met the four attributes. Thus, the first step was to ensure that green start-ups indeed complied with those attributes and for this; the answers to four questions were required:

(i) Balanced Green Investment (BGI): does your company seek a balance between economic and environmental aims?

(ii) Innovation and Competitiveness (IC): has your company implemented innovative green ideas and practices in the market?

(iii) Environmental Impact (EI): has your company implemented green technologies to prevent contributing to the environmental degradation of nature?

Table 1: Online environmental startups in Spain.

Sector	Population	Sample	Sector	Population	Sample
Agriculture	2	1	Fashion	3	3
Art	4	4	Finance	4	3
Automotive	4	4	Food	9	7
Clean Technology	15	11	Funeral Industry	1	1
Construction	3	3	Health and Wellness	10	9
Consulting	11	12	Health Care	2	2
Cosmetic	4	4	Human Resources	1	1
Design	10	9	Information Technology	15	10
Digital Marketing	2	2	Internet of things	5	4
E-commerce	8	8	Life Sciences	1	1
Education	6	5	Logistic	4	3
Electric bicycle	4	3	Nature	7	6
Energy	14	10	Restaurants	11	7
Entertainment	4	2	Travel	16	9
Events	3	2	Urban development	3	2
Farming	3	2	Water	3	2
Total	97	82	Total	95	70
Total Population	192		Total Sample	152	

(iv) Green Opportunities (GO): does your company identify the Green Opportunities that the market offers?

A brief presentation of the aims of the research was also sent to them in the context of the Spanish Corporate Governance Code [23] in what concerns corporate compliance. Two hundred and twelve start-ups answered and asserted complying with the four attributes. The survey was sent to them, and one hundred fifty-two companies returned the survey with their answers. The team then emailed the compliance officers and several heads of departments between June and August 2018 requesting they answer the survey. The structure and distribution of the population under study and the sample are explained in Table 1.

3.2. Surveys. According to the literature review, a survey was drafted to measure green entrepreneurs' attitudes towards nature empirically. Twenty statements were completed.

Two focus groups held online meetings via Skype to validate the survey. Skype was used to overcome the distance between participants. Nine green entrepreneurs from different areas of Spain were involved. Twenty original questions were discussed during a period of two hours in each meeting. The survey was amended as a result. Five of the original items were deleted and two added. With this information, an experiential survey was conducted to validate the proposed survey. Six interviews were made to contrast the clarity of the questions. Three questions were further modified after the pretest. The final survey is shown in Table 2.

The items were analysed through the ten-point Likert scale to indicate the degree of importance of the factors (1 = "fully disagree" to 10 = "fully agree") (Allen and Seaman, 2007).

3.3. Model and Data Analysis Process. SEM-PLS Path Modelling was used to ascertain the model and obtain results [62]. SEM not only enables examining the relationships between observable indicators and constructs statistically [63], but also works with composite model latent variables [62]. The methods can be used for explanatory and predictive research as well as complex models.The green start-up model is composed of three endogenous constructs, the Environmental Value Model (EVM), Green Attitude towards Nature (GAN), and Environmental Corporate Compliance (ECC), and four exogenous ones, Balanced Green Investment (BGI), Innovation and Competitiveness (IC), Environmental Impact (EI), and Green Opportunities (GO), (see Figure 2).

Seven hypotheses were analysed in the study:

H_1: Environmental Corporate Compliance (ECC) positively influences Green Attitudes towards Nature (GAN) among entrepreneurs.

H_2: Environmental Corporate Compliance (ECC) positively influences the Environmental Value Model (EVM).

H_3: the Environmental Value Model (EVM) positively influences Green Attitudes towards Nature (GAN).

H_4: the Environmental Value Model (EVM) positively influences Balanced Green Investment (BGI).

H_5: the Environmental Value Model (EVM) positively influences Innovation and Competitiveness (IC).

H_6: the Environmental Value Model (EVM) positively influences Environmental Impact (EI).

H_7: the Environmental Value Model (EVM) positively influences Green Opportunities (GO)

Table 2: Latent variables and the elaborated questionnaire.

Latent variables	Questions
GAN: Green Attitudes towards Nature	Do you think it is important for your company to be aware of being an appropriate steward of natural resources (GAN_1)? Do you think it is important for your company to reduce hazardous emissions or toxic materials to improve health and safety at work (GAN_2)?
ECC: Environmental Corporate Compliance	Do you think it is important to implement the prevention model to supervise the compliance of green innovations and procedures and services with the law (ECC_1)? Do you think it is important to undertake protocols and training procedures for staff members to apply these surveillance and organisational measures (ECC_2)? Do you think it is important to establish a disciplinary system to prevent non-compliance with the law (ECC_3)?
The Environmental Value Model (EVM).	Do you think it is important to build entrepreneurial models to create environmental values based on sustainable principles (EVM_1)? Do you think it is important to describe, analyse and communicate such values in your company (EVM_2)?
Balanced Green Investment(BGI)	Do you think it is important to seek a balance between profit, nature and people through sustainable management to protect nature (BGI_1)? Do you think it is important to reach environmental, economic and social standards to respect future generations in an innovation context (BGI_2)?
Environmental Impact (EI)	Do you think it is important to produce green technologies to prevent environmental degradation on the planet (EI_1)? Do you think it is important to reduce the consumption of toxic substances and harmful emissions to prevent damages to the health and safety of employees at work (EI_2)?
Green Opportunities (GO)	Do you think it is important to develop green skills among employees to build opportunities to expand environmental values in the company (GO_1)? Do you think it is important to implement a new generation of manufacturing processes to reduce pollution in production (GO_2)? Do you think it is important to implement eco-friendly technologies to overcome the market failures (GO_3)?
Innovation and Competitiveness (IC)	Do you think it is important to propose new ideas and support creative processes (IC_1)? Do you think it is important to become a pioneer in green innovation ideas and practices (IC_2)? Do you think it is important to respond faster than your competitors to the needs of customers (IC_3)?

4. Results

4.1. Measurement Model. Reliability and validity are the first two conditions used to assess the model. The process can be structured in four steps: (1) individual item reliability, (2) construct reliability, (3) convergent validity, and (4) discriminant validity. First, individual reliability is measured through the load (λ) of each item. The minimum level established for acceptance as part of the construct is typically $\lambda >= 0.707$ [64]. This condition was validated in the model (see Figure 3).

Composite reliability (CR) was applied to test the consistency of the constructs. This evaluation measures the rigour with which these elements measure the same latent variable [65].

Cronbach's alpha index also determines the consistency of the model for every latent variable. Values higher than 0.7 are typically accepted [66]. Table 3 shows that the reliability of each construct was accepted.

AVE measures the convergent validity, the acceptable limit of which is 0.5 or higher. This indicator provides information about the level of convergence of the constructs with their indicators [67]. Table 3 also shows that the values meet the criteria. Rho_A was also measured. Results exceed the value of 0.7 [68].

Table 4 explains the correlations between the constructs on the left. A construct should share more variance with its indicators than with other latent variables in the model [69]. However, Henseler et al. [70] did detect a lack of discriminant validity in this respect. Ratio Heterotrait-monotrait (HTMT) provides a better approach to this indicator. The results obtained in this sense are on the right. The HTMT ratio for each pair of factors is <0.90.

4.2. Structural Model Analyses. The structural model of assessment proposed is explained in Table 5. The general criterion used to evaluate the structural model is the coefficient of determination (R-squared). R-squared analyses the proportion of variance (in percentage) in the exogenous variable that can be conveyed by the endogenous variable. The R-squared value can be expressed from 0 to 1. Values close to 1 define the predictive accuracy. Chin [71] proposed a rule of thumb for acceptable R-squared with 0.67, 0.33, and 0.19. They are defined as substantial, moderate, and weak predictive power, respectively.

Table 3: Cronbach Alpha, rho_A, Composite Reliability, and AVE.

	Cronbach's Alpha	rho_A	Composite Reliability	Average Variance Extracted (AVE)
ECC	0,874	0,895	0,876	0,706
EI	0,810	0,812	0,810	0,681
EVM	0,853	0,854	0,854	0,745
GAN	0,820	0,826	0,822	0,698
BGI	0,817	0,827	0,820	0,696
GO	0,849	0,853	0,848	0,651
IC	0,864	0,869	0,865	0,683

Table 4: Measurement model. Discriminant validity.

	Fornell-Larcker Criterion							Heterotrait-monotrait ratio (HTMT)						
	ECC	EI	EVM	GAN	BGI	GO	IC	ECC	EI	EVM	GAN	BGI	GO	IC
ECC	0,840													
EI	0,644	0,825						0,655						
EVM	0,697	0,730	0,863					0,694	0,729					
GAN	0,793	0,557	0,830	0,836				0,793	0,561	0,832				
BGI	0,703	0,406	0,780	0,745	0,834			0,712	0,408	0,783	0,760			
GO	0,456	0,491	0,773	0,557	0,571	0,807		0,452	0,494	0,770	0,562	0,576		
IC	0,499	0,268	0,597	0,652	0,550	0,552	0,826	0,496	0,269	0,599	0,652	0,557	0,556	

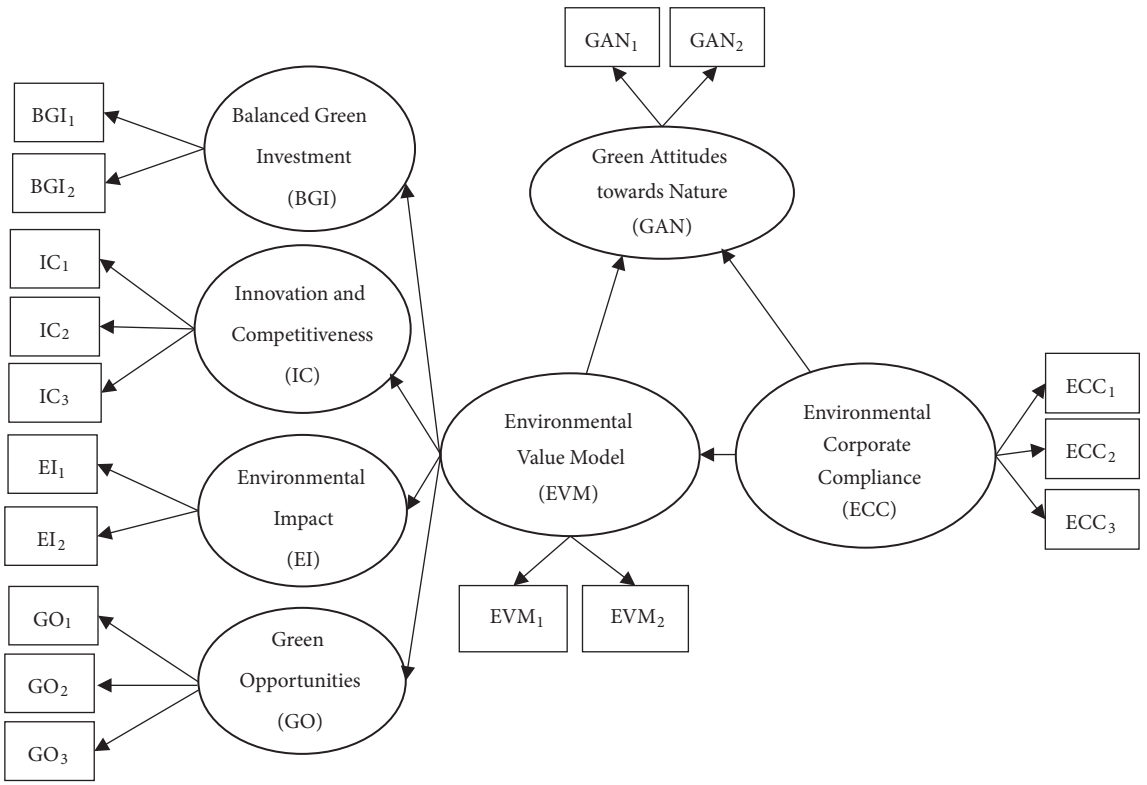


Figure 2: Conceptual scheme of the structural equation model utilized. BAN: Green Attitudes towards Nature. ECC: Environmental Corporate Compliance. EVM: Environmental Value Model. BGI: Balanced Green Investment. EI: environmental impact. IC: innovation and competitiveness. GO: Green Opportunities. EP: Educational Process.

Table 5: Comparison of hypothesis.

Hypotheses	Effect	Path coefficient (β)	Confident Interval (2.5%)	Confident Interval (95%)	t-statistic (β/STDEV)	p-Value	Supported
H1	ECC⟶ GAN	0.417	0,096	0,665	5.116	0,009	Yes ∗∗
H2	ECC ⟶ EVM	0.697	0,551	0,825	2.905	0,000	Yes ∗ ∗ ∗
H3	EVM ⟶ GAN	0.539	0,288	0,864	4.760	0,001	Yes ∗∗
H4	EVM ⟶ BGI	0.780	0,673	0,878	6.375	0,000	Yes ∗ ∗ ∗
H5	EVM ⟶ IC	0.597	0,462	0,727	2.737	0,000	Yes ∗∗
H6	EVM ⟶ EI	0.730	0,548	0,873	8.395	0,000	Yes ∗ ∗ ∗
H7	EVM ⟶ GO	0.773	0,692	0,850	4.476	0,000	Yes ∗∗

Note. For n = 5000 subsamples, for t-distribution (499) Students in single queue: ∗ p < 0.05 (t (0.05;499) = 1.64791345); ∗∗ p < 0.01 (t(0.01;499) = 2.333843952); ∗ ∗ ∗ p < 0.001 (t(0.001;499) = 3.106644601).

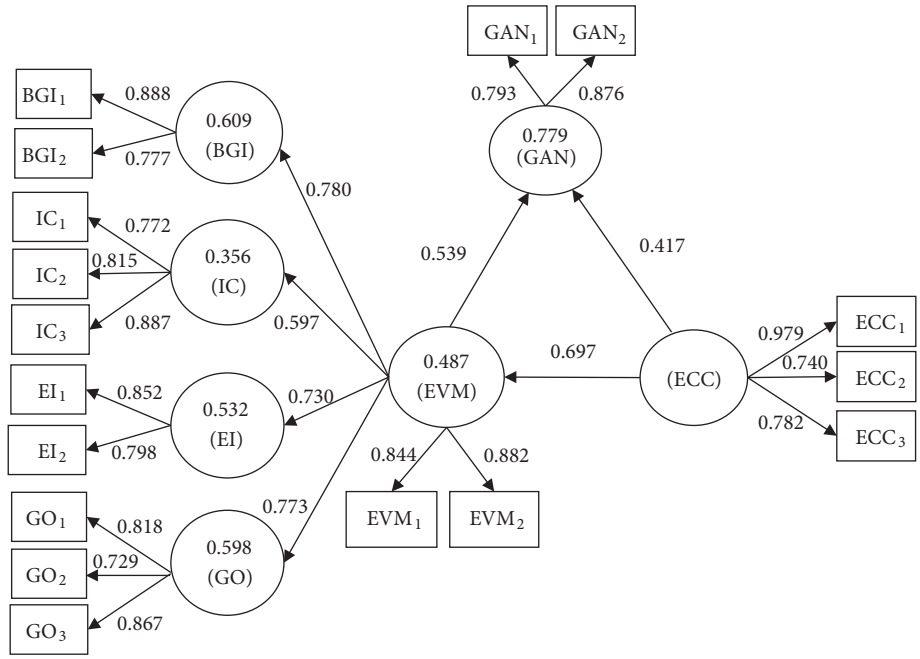


Figure 3: Individual item reliability.

Our model presents constructs with high predictive value (R-squared⟶GAN=0.776) and moderate predictive value (R-squared⟶BGI=0.606; R-squared⟶GO=0.595; R-squared⟶EI=0.529; R-squared⟶IC=0.352) and weak R-squared⟶EVM=0.268). Therefore, the evidence shows that this model is applicable for green start-ups when developing Green Attitudes towards Nature due to its strong predictive power and explanatory capacity.

Table 5 also shows the hypothesis testing using 5000 bootstrapped resamples. After analysing the path coefficients, there are grounds to assert that all hypothesised relationships are significant at 99.9% confidence levels, except two of them; the first one is ECC⟶GAN, β=0.417, Statistical T=5.116, and the second one is EVM⟶GAN, β=0.539, Statistical T=4.760. These are supported with a 99% confidence level.

As part of the assessment of the structured model, SRMS also needs to be measured to analyse the good fit of the model. In the research, SRMR is 0.075, which is less than 0.08, as Hu and Bentler [72] had expressly indicated.

Blindfolding is also assessed within the model. It measures the predictive capacity of the model through the Stone-Geisser test (Q^2) [73, 74]. The result revealed that the model is predictive ($Q^2 = 0.474$) since $Q^2 > 0$.

4.3. Results for Unobserved Heterogeneity. The unobserved heterogeneity can to be analysed with different PLS segmentation methods [75, 76]. We select FIMIX-PLS and PLS prediction-oriented segmentation (PLS-POS) methodology for two reasons [76]. First, according to the evaluation of these methods, Sarstedt [75] concludes that FIMIX-PLS is

Table 6: Indices FIT. Criteria for model choice.

	K=2	K=3	K=4	K=5
AIC (criterio de información de Akaike)	1.948,641	1.907,489	1.884,002	**1.864,526**
AIC3 (modificado de AIC con Factor 3)	1.975,641	1.948,489	1.939,002	**1.933,526**
AIC4 (modificado de AIC con Factor 4)	2.002,641	**1.989,489**	1.994,002	2.002,526
BIC (criterio de información Bayesiano)	**2.030,285**	2.031,468	2.050,315	2.073,173
CAIC (AIC consistente)	**2.057,285**	2.072,468	2.105,315	2.142,173
MDL5 (longitud de descripción mínima con Factor 5)	**2.572,865**	2.855,384	3.155,569	3.459,765
LnL (LogLikelihood)	-947,320	-912,744	-887,001	-863,263
EN (estadístico de entropía (normalizado))	0,672	**0,724**	0,722	**0,767**

viewed as the proper commonly used approach to capture heterogeneity in PLS Path Modelling. Second, PLS-POS informs about nonobserved heterogeneity in the structural model as well as the constructs measures, with both formative and reflective models [76].

In order to achieve a major capacity of prediction PLS-POS provides ongoing improvements of the objective targeted. Due to the hill-climbing approach, iterations of the algorithm might generate an intermediate solution, not good enough to be validated. For this reason, it is important to run the application of PLS-POS with different starting segmentations [76]. In our case, we have applied the PLS-POS algorithm with 5, 4, 3, and 2 partitions.

Regarding Becker et al. [76], the bias from using either of the two methods (FIMIX-PLS or PLS-POS) is much lower than that obtained from analysing the overall sample without uncovering heterogeneity [76].

In addition, FIMIX-PLS is understood as better methods for reducing biases in parameters estimates and avoiding inferential. Becker et al. [76] find an exception in low structural model heterogeneity and high formative measurement model heterogeneity. Regarding this condition, FIMIX-PLS generate more biased results than those resulting from ignoring heterogeneity [76].

In our case, we do not find unobserved heterogeneity with PLS-POS algorithm because the results indicate one group too small for the study in all iterations.

As a result, we have used de FIMIX-PLS. This methodology, considers the possibility of acceptance in any segment observed. The observations are adapted depending on the number of segments. Through the linear regression functions a group of potential segments is given. Every case is attached to the segment with the highest probability. FIMIX methodology was applied to structure the sample into several segments. Selecting the convenient number of segments was the first problem found. It is usual to repeat the FIMIX-PLS method with successive numbers of latent classes [77]. In this case, taking into account the sample size n = 152, the calculation was made for k=2, k=3, k=4, and k=5. Different information criteria offered by the FIT indices were used to compare the results obtained. The controlled AIC (CAIC), Akaike (AIC), the standardized entropy statistic (EN), and the Bayesian information criterion (BIC) were compared.

The study is implemented in four stages: in the first place, FIMIX provides the number of best segments. Hence, the construct that confirms these segments is found. As a result, the segments and the model are estimated. Table 6 shows the results provided for the FIT indices.

Firstly, in order to find the number of samplings, it can be organised into the FIMIX test applied. The algorithm for the size of the sample so as to be able to use PLS-SEM with 10 repetitions was constructed. As a result, the new composition was carried out using the expected maximization algorithm (EM). The EM algorithm switches between maximization step (M) and performing an expectation step (E) [78]. Step E assesses the accepted estimation of the variables. Step M measures the parameters by making as big as possible the logarithmic registration likelihood obtained in step E. Steps E and M are used continuously until the results are balanced. The equilibrium is attained when no essential progress in the values is attained.

The results after running FIMIX with different numbers of partitions are shown in Table 6. As the number of segments was unknown beforehand, the different segment numbers were contrasted in terms of appropriateness and statistical analysis [79, 80].

Trial and error information can be conveyed within the EN and information criteria due to their sensitiveness to the features of the model and their data. The data-based approach gave a rough guide of the number of segments [78].

As a result, the criteria applied were assessed. The ability to accomplish aim of the information criteria in FIMIX-PLS was studied for a broad range of data groups [81]. Their outputs indicated that researchers should take into account AIC 3 and CAIC. While these two criteria express the equivalent number of segments, the results possibly show the convenient number of segments. Table 6 shows that in our study these outputs do not indicate the same number of segments, so AIC was utilized with factor 4 (AIC 4, [82]). This index usually behaves appropriately. The same number of segments was shown in the case study (see Table 6), which was k=3. Then, it is understood to be a strong miscalculation, even though MDL5 expressed the minimum number of segments k+1. In this case it would imply 3 [78]. The regulated entropy statistic (EN) was one of the measurements of entropy which was also esteemed [83]. EN applied the likelihood that an observation is addressed to a segment to express if the separation is trustworthy or not. The greater the chance of being included to a segment is for an assessment, the higher segment relationship is. The EN index varies between 0 and 1. The greatest values express the better quality segmentation. Other researches in which EN

Table 7: Relative segment sizes.

K	Segment 1	Segment 2	Segment 3	Segment 4	Segment 5
2	0,537	0,463			
3	**0,384**	**0,310**	**0,307**		
4	0,342	0,325	0,258	0,074	
5	0,462	0,222	0,138	0,131	0,047

Table 8: Path coefficients for the general model and for the three segments.

Path Coefficients	Global model	k=1 (38.4%) n=54	k=2 (31%) n=51	k=3 (30.7%) n=47	MGA k1 vs k2	MGA k1 vs k3	MGA k2 vs k3
ECC⟶EVM	0.620∗ ∗ ∗	0.274∗	0.742∗ ∗ ∗	0.932∗ ∗ ∗	0.468 n.s.	0.659 n.s.	0.190 n.s.
ECC⟶GAN	0.340∗ ∗ ∗	0.363∗∗	-0.033 n.s.	0.015 n.s.	0.397∗∗	0.348∗	0.049 n.s.
EVM⟶EI	0.606∗ ∗ ∗	0.456∗ ∗ ∗	0.548∗ ∗ ∗	0.794∗ ∗ ∗	0.091 n.s.	0.338 n.s.	0.246 n.s.
EVM⟶GAN	0.549∗ ∗ ∗	0.311∗	0.986∗ ∗ ∗	0.978∗ ∗ ∗	0.675 n.s.	0.667 n.s.	0.008 n.s.
EVM⟶GBI	0.692∗ ∗ ∗	0.567∗ ∗ ∗	0.956∗ ∗ ∗	0.541∗ ∗ ∗	0.389 n.s.	0.026 n.s.	0.415∗ ∗ ∗
EVM⟶GO	0.649∗ ∗ ∗	0.550∗ ∗ ∗	0.728∗ ∗ ∗	0.570∗ ∗ ∗	0.178 n.s.	0.021 n.s.	0.157∗
EVM⟶IC	0.537∗ ∗ ∗	0.095 n.s.	0.943∗ ∗ ∗	0.550∗ ∗ ∗	0.848 n.s.	0.455 n.s.	0.393∗ ∗ ∗

Note: n.s., not supported; ∗ $p<0.05$; ∗∗ $p<0.01$; ∗ ∗ ∗ $p<0.001$.

values are above 0.50 which offer easy understanding of the data into the chosen number of segments are also appointed [84, 85]. Table 6 shows that all the segmentations conveyed EN values>0.50, even though the greatest value is achieved for k = 5 with EN=0.767 for k=5 and EN=0.724 for k=3.

Accordingly, the number of best segments was k=3. FIMIX-PLS indicates, then, the number of segments, due to the lowest size of the segmentation, which in this case is 30.7%. Table 7 shows that, for the k=3 solution and a sample n=152, segment partitioning is 38.4% (54), 31% (51), and 30.7% (47) [72, 86]. In spite of the percentages, the segment sizes are significant, so they are enough to use PLS. Due to the covariance, the sample size can be substantially lowest in PLS than in SEM [77]. It means that it might be more variables than observations, which imply data missing can be obtained [87, 88]. Similarly, in similar cases other authors have indicated the reduced size of the sample [89] whose minimum might attain the number of 20 in PLS [90].

The process of the FIMIX-PLS strategy is completed with these analyses. On the other hand, other researches suggest testing to see if the numerical variation between the path coefficients of the segment is also significantly distinctive by using multigroup analysis (see Table 8). Several approaches for multigroup analysis were found in document research, which are discussed in more detail by Sarstedt et al. [81] and Hair et al. [66]. The use of the permutation approach was suggested by Hair et al. [78], which was also used in the SmartPLS 3 software.

However, before making the interpretation of the multigroup analysis outcomes researchers must ensure that the measurement models are invariable in the groups. Once the measurement invariance (MICOM) was checked [79], a multigroup analysis (MGA) was carried out to find if there were any significant discrepancies between the segments. The results are shown in the three right hand side columns of Table 9.

As proven by the results obtained from nonparametric testing, the multigroup PLS-MGA analysis confirmed the parametric tests and also found significant discrepancy between segments 2 and 3.

There is divergence between the first and second segments, but only k=2 and K=3 in EVMGBI, EVMGO, and EVMIC show a significant difference.

Table 9 shows the validity of the segment measurement model and its explanatory capacity using R-squared and classified by segment. The values of k=2 for CR and AVE are shown to be below the limits.

4.3.1. Assessment of the Predictive Validity. PLS can be used for both explanatory and predictive research. In other words, the model has the ability to predict the current and future observations. Predictive validity shows that the measurements for several constructs can predict a dependent latent variable, as, in our case, Green Attitudes towards Nature (GAN). The prediction outside the sample, known as predictive validity, was assessed using cross-validation with maintained samples. This research used the approach recommended by Shmueli et al. [91].

By using other authors' research [92–94], the PLS predict algorithm in the updated SmartPLS software version 3.2.7 was used. The results for k-fold cross prediction errors and the essence of prediction errors. It was expresses through the mean absolute error (MAE) and root mean square error (RMSE). Then, the expected achievement of the PLS model for latent variables and indicators is assessed. The following criterion expressed by the SmartPLS team was applied to appraise the expected achievement of the model [91–94].

(1)The Q^2 value in PLS predict: the miscalculations of the PLS model with the mean future projection are compared.

The PLS-SEM's miscalculation data can be lower than the prediction error of simply using mean values; then the Q^2 value is positive. Therefore, the PLS-SEM model provided

Table 9: Reliability measurements for the general model and for the three segments.

	Global model			k=1 (38.4%)			k=2 (31%)			K=3 (30.7%)		
	CR	AVE	R-squared	CR	AVE	R-squared	CR	AVE	R-squared	CR	AVE	R-squared
ECC	0.922	0.797	-	0.947	0.857	-	0.791	0.561	-	0.912	0.775	-
EI	0.913	0.840	0.368	0.935	0.877	0.208	0.857	0.750	0.300	0.886	0.795	0.631
EVM	0.920	0.794	0.384	0.861	0.675	0.075	0.918	0.789	0.551	0.959	0.885	0.869
GAN	0.917	0.847	0.648	0.908	0.831	0.290	0.861	0.756	0.924	0.940	0.886	0.985
GBI	0.916	0.845	0.479	0.851	0.741	0.322	0.893	0.807	0.914	0.945	0.895	0.293
GO	0.909	0.768	0.421	0.878	0.707	0.302	0.861	0.676	0.529	0.929	0.814	0.325
IC	0.917	0.786	0.288	0.904	0.759	0.009	0.872	0.694	0.899	0.938	0.835	0.303

Table 10: Summary of dependent variable prediction.

Construct GAN	RMSE	MAE	Q^2
Complete sample	0.591	0.425	0.429
Segment 1	0.486	0.376	0.120
Segment 2	0.619	0.468	0.515
Segment 3	0.723	0.490	0.761

convenient predictive performance, which is the case in the two subsamples of segments 2 and 3 (see Table 10) in the dependent construct Green Attitude towards Nature (GAN) (Table 10). Then, the prediction results were achieved.

(2) The linear regression model (LM) approach: a regression of the exogenous indicators in every endogenous indicator was performed. Then, better prediction errors can be achieved when this comparison is considered. This can be seen when the MAE and RMSE values are smaller than those of the LM model. If this occurs, predictions can be made. This methodology is only used for indicators. As shown in Table 10, the MAE and RMSE values were mainly negative. It expressed excellent predictive power.

It is also contrasted the predictions the real composite scores within the sample and outside the sample with [91]. With this intention, the research by Danks, Ray, and Shmueli [95] was applied.

By using this methodology, the measure was applied for the Green Attitudes towards Nature (GAN) construct: RMSE for the complete sample (see Table 10) was 0.591 and had a higher value in segment 3 (0.723, difference=0.132) and lower values in segment 1 (0.486, difference=0.105) and segment 2 (0.619, difference=0.028). The complex values are normalized in which the value of mean is 0 and variance 1. RMSE expressed the standard deviation measure. Since the difference in RMSE is not considerable, excess capacity is not a problem for this study.

In relation to Q^2, the following metrics were found for the GAN construct: RMSE for the complete sample (see Table 10) was 0.429 and had a higher value in segment 3 (0.761, difference=0.332) and lower values in segment 1 (0.120, difference=0.309) and segment 2 (0.515, difference=0.086).

The density diagrams of the residues within the sample and outside the sample are provided in Figure 4.

Due to the result of the different analyses, this research found enough evidence to accept the predictive validity of the

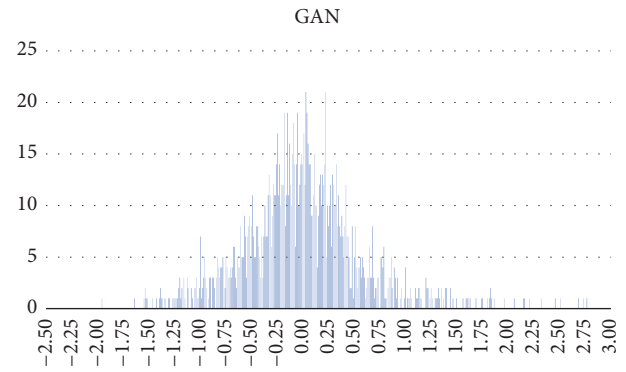


Figure 4: Residue density within the sample and outside the sample.

model expressed. As a result, the model conveys appropriately the intention to apply in further samples. They are quite distinctive from the data used to check the theoretical model [96].

4.4. Considerations for the Management of Internet Search Engines (IPMA). According to research that studied data heterogeneity [77, 96], the IPMA-PLS technique was used to find more precise recommendations for marketing of Internet search engines. IPMA is a framework study that uses matrices that enable combining the average value score for "performance" with the estimation "importance" in PLS-SEM's total effects [77, 97, 98]. The outcomes are shown in an importance-performance chart of four fields [77, 99].

According to Groß [97] the analysis of the four quadrants is shown in the chart (Figure 5). They are expressed consequently in the following points:

(i) Quadrant I conveys acceptance attributes that are much more valued for performance and importance.

(ii) Quadrant II explains acceptance attributes of high importance but small performance. It must be developed.

(iii) Quadrant III considers acceptance attributes that have reduced importance and performance.

(iv) Quadrant IV expresses acceptance attributes with great performance index, but small equitable importance.

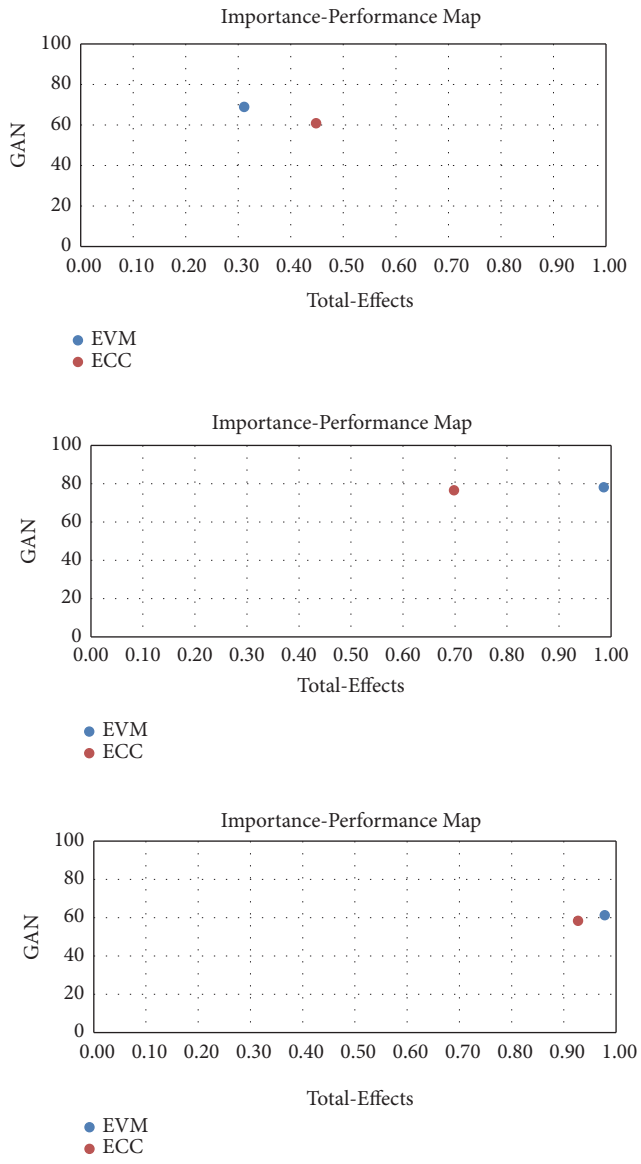


Figure 5: Importance-performance maps for k=1, k=2, and k=3.

The results are unequal for each segment (Figure 5). For users belonging to k=1, all constructs are low and provide performance <50, showing that ECC and EVM obtained a score of 0.45 and 0.31.

Finally, the results for k=3 show a different situation. These entrepreneurs valued importance >90, showing that ECC and EVM is the most valued and obtained a score of 0.93 and 0.98.

5. Discussion

5.1. Theoretical Implications. Spanish green start-ups have increasingly developed better green attitudes in recent decades. The new environmental regulatory system has been the driving force behind several changes in the way green start-ups must comply with the law. On the one hand, the Spanish regulatory compliance has established a new legal framework for companies. On the other hand, it proposed the implementation of an organisational model to prevent companies committing unexpected unlawful behaviour.

The FIMIX-PLS analysis has been split into two groups. Coincidentally, the size of the segments and the coefficient of determination R-squared as the FIT indices are divided into two samples groups as well.

Nonparametric unobserved segmentation in complex systems has provided us enough information to compare the statistical results. To present the complete analyses within FIMIX-PLS method multigroup and permutation approach have been taken into account. Results show that segments have presented differences using the multigroup analysis (MGA), namely, between the first and second segments: EVM⟶GBI, EVM⟶GO, and EVM⟶IC.

To revise the validity of the segments measured SmartPLS software version 3.2.7 was used. This statistical package was used to set up the predictive performance of the model. The prediction errors of the PLS model were compared with the simple mean predictions. In the two subsamples of segments 2 and 3 in the dependent construct GAN the prediction error of the PLS-SEM results was less than the prediction error of simply using the mean values. It means that the model provides acceptable prediction results. Similarly, the linear regression model (LM) approach was studied in order to get better prediction errors. It is achieved when the value of RMSE and MAE are lower than those of the LM model. As the results have shown, the indicators show the valued of RMSE and MAE were mostly negative. It can be deduced that those values provide a strong predictive power.

Such findings are manifested through the significance of the path coefficients, particularly, among the influence of Environmental Corporate Compliance in the Environmental Value Model (H_2: ECC⟶EVM; β = 0.697, T-statistic = 2.905) and in the Green Attitudes towards Nature (H_1: ECC⟶GAN; β = 0.417, T-statistic = 5.116). As a result, regulatory compliance not only prevents companies from committing unlawful environmental actions [39] but also from running the risk of any negative implications for nature [4].

The T-statistic in both path coefficients shows little difference. Nevertheless, it is safe to say that the new regulation has a higher effect on improving attitudes towards nature than the Environmental Value Model.

At the empirical level, the new regulation has a significant effect on Spanish green start-ups as a way of improving attitudes towards nature and its Environmental Value Model. In other words, the new Criminal Code raises awareness on monitoring a voluntary environmental surveillance process based on the rules of the organisation [36]. Simultaneously, start-ups are willing to update the environmental standards that allow them to comply with the law by developing a preventive system to deter unlawful decisions taken by its managers [34].

Start-ups should, therefore, implement effective monitoring systems, including organisational monitoring, to avoid unexpected criminal sanctions [38] and keep the compliance model up to date when relevant infractions are revealed [40].

The findings also show that the hypothesised relationships between Environmental Value Model and the four attributes (Balanced Green Investment, Environmental Impact of green entrepreneurs, Innovation and Competitiveness and Green Opportunities) are indeed significant. Based on the results obtained, each attribute creates a real environmental value for start-ups. The environmental value makes them more innovative and competitive (H_5: EVM⟶IC; $\beta = 0.597$, ≤ 0.001) and green-oriented towards market opportunities (H_7: EVM ⟶ GO; $\beta = 0.773$, p value ≤ 0.001) by balancing economy and ecology (H_4: EVM ⟶ BGI; $\beta = 0.780$, p value ≤ 0.001). All these factors contribute towards generating a positive impact in nature (H_6: EVM ⟶ EI; $\beta = 0.730$, p value ≤ 0.001).

5.2. Practical Implications. Practical implications can be drawn from the findings not only for green start-ups but also for public and legal authorities. The first implication is related to the need for empirical studies to test the recent approval of the Spanish regulatory compliance as well as nurture their legal and public decisions. The second implication is directly related to the findings. Results have offered an unexpected picture of what is commonly understood about coercive regulations. Results show that the new regulatory compliance system has emerged as the critical factor to foster green start-ups' respect for nature. The new regulation positively influences start-ups' improvement towards nature (H1: ECC⟶GAN; =0.417, T-statistic=5.116) and has a positive effect on the Environmental Value Model (H2: ECC⟶EVM; = 0.697, T-statistic=2.905). The new environmental law, therefore, plays a key role to explain how these entrepreneurs respect nature, whether by setting up a department to supervise the compliance of green innovations [34] or by undertaking training protocols to implement surveillance norms in the companies [36]. The research also contributes, as the third implication, towards protecting companies from unlawful actions and corruption by developing organisational indicators. In other words, corruption, which is supposed to be negatively associated with private investment and growth, has become an outstanding setback and international phenomenon [98]. This explains the current trend to avert bribery, extortion, or fraud by developing international efforts to address this dilemma from the legal and organisational perspective.

The fourth implication alludes to the model designed. As the results have shown, H4, H5, H6, and H7 are significant. In other words, the four attributes appropriately reflect the Environmental Value Model. On top of that, a balanced economy and ecology are considered the main attributes (H4: EVM⟶BGI =0.780, p value 0.001). Finally, green start-ups are willing to concentrate efforts on environmental and managerial solutions by discovering new Green Opportunities to defend nature. This trend, devised in green entrepreneurs, might positively influence the future of start-ups, as confirmed by the predictive value of the model ($Q^2 = 0.474$).

6. Conclusion

Two relevant contributions have been made to the literature review. The first one has been the methodology it has been applied. This is so not only because it has been used non-parametric unobserved segmentation in complex systems, but also because it gives you another approach about data can be organised to provide statistical results.

The second contribution is that the outcome of this study provides relevant theoretical and practical implications to be incorporated into the literature review. Findings shed light on not only the way green start-ups comply with the recent Spanish Criminal Code, but also how they develop green attitudes towards nature to avoid threats to nature.

This research highlights three aspects regarding the effects of the new regulatory compliance system in Spain: (1) the implications of the new regulation on green start-ups by imposing coercive rules and designing voluntary surveillance measures; (2) the significance of the Environmental Value Model reflected in four green attributes; and (3) the influence of the Environmental Value Model and the new regulation in green start-ups' attitudes towards nature.

The findings provided a robust explanatory capacity of the complete model (R-squared GAN=0.779). In other words, all the constructs explain 77.9% of the variance of start-ups' Green Attitudes towards Nature.

Theoretical and practical implications can also be learnt from this robust result. Even though the literature about green companies is vast, very little has been published about the empirical impact of environmental regulation on green start-ups.

This study helps to shed light on this aspect. Due to the gradual increase in the number of start-ups in the last decades, these findings can be used by public authorities to impose legal requirements and encourage companies to develop surveillance measures through their compliance officers.

To be more precise, we recommend the authorities focus on promoting organisational measures, by rewarding companies that establish protocols or training procedures, and set up surveillance standards within the company. This measure can also be undertaken by developing interconnected mechanisms between public authorities and start-ups based on cooperative rules as a priority to improve the employees' attitudes towards nature. Three limitations of the study must also be appropriately conveyed.

First. A new list of Spanish green start-ups has been proposed in the study by selecting four attributes from the literature review.

Second. Not all start-ups were aware of the recent publication of the Spanish Corporate Compliance Criminal Code let alone implemented it.

Third. The research was based on collecting data on how green start-ups perceived the new regulation and their influence on the model presented. Unfortunately, the research did not offer further information about the design process of organisational standards to prevent unlawful decisions. In other words, a practical and theoretical approach to that phenomenon might yield different results.

Concerning future research lines, it would be advisable to work on these limitations to achieve a more accurate

approach to the current research. Moreover, the research outcomes obtained can be compared with nonenvironmental start-ups, to see how they see nature, as a critical element to be respected by every company to preserve the environment for our future generations.

The research concludes by stating that the three relationships were successfully tested and contrasted by the model. The new Spanish regulatory compliance has provided unexpected results that could contribute to improve the legal decisions taken by public authorities.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

There are no conflicts of interest.

Authors' Contributions

Rafael Robina-Ramírez wrote the theoretical and empirical part. Antonio Fernández-Portillo addressed the FIMIX results. Juan Carlos Díaz-Casero has revised the paper.

Acknowledgments

INTERRA (Research Institute for Sustainable Territorial Development) has collaborated in the research. We have not received any funds.

References

[1] European Commission, EU, SBA Fact Sheet - European Commission - Europa EU. Ref. Ares (2018) 2717562 - 25/05/2018. 2018. Retrieve: https://ec.europa.eu/docsroom/documents/29489/attachments/27/translations/en/renditions/pdf.

[2] G. H. Brundtland, "Brundtland report," Our Common Future. Comissão Mundial, 1987.

[3] R. Barkemeyer, "Beyond compliance - below expectations? CSR in the context of international development," *Business Ethics: A European Review*, vol. 18, no. 3, pp. 273–289, 2009.

[4] B. Cohen and M. I. Winn, "Market imperfections, opportunity and sustainable entrepreneurship," *Journal of Business Venturing*, vol. 22, no. 1, pp. 29–49, 2007.

[5] J. Randjelovic, A. R. O'Rourke, and R. J. Orsato, "The emergence of green venture capital," *Business Strategy and the Environment*, vol. 12, no. 4, pp. 240–253, 2003.

[6] L. U. Hendrickson and D. B. Tuttle, "Dynamic management of the environmental enterprise: a qualitative analysis," *Journal of Organizational Change Management*, vol. 10, no. 4, pp. 363–382, 1997.

[7] J. D. Ivanko, *ECOpreneuring Putting Purpose and the Planet Before Profits*, New Society Publishers, Gabriola Island, Canada, 2008.

[8] C. Leadbeater, *Social Enterprise and Social Innovation: Strategies for the Next Ten Years. A Social Enterprise Think Piece for the Office of the Third Sector*, Cabinet Office Office of the Third Sector, London, UK, 2007.

[9] S. Schaltegger, "A framework for ecopreneurship: Leading bioneers and environmental managers to ecopreneurship," *Greener Management International*, no. 38, pp. 45–58, 2002.

[10] T. J. Dean and J. S. McMullen, "Toward a theory of sustainable entrepreneurship: Reducing environmental degradation through entrepreneurial action," *Journal of Business Venturing*, vol. 22, no. 1, pp. 50–76, 2007.

[11] S. Schaltegger, "A framework and typology of ecopreneurship: leading bioneers and environmental managers to ecopreneurship," in *Making Ecopreneurs*, M. Schaper, Ed., pp. 95–114, Routledge, London, UK, 2016.

[12] Y. Li, Z. Wei, and Y. Liu, "Strategic orientations, knowledge acquisition, and firm performance: the perspective of the vendor in cross-border outsourcing," *Journal of Management Studies*, vol. 47, no. 8, pp. 1457–1482, 2010.

[13] E. E. Walley and D. W. Taylor, "Opportunists, champions, Mavericks...? a typology of green entrepreneurs," *Greener Management International*, no. 38, pp. 31–43, 2002.

[14] A. R. Anderson, "Cultivating the garden of eden: environmental entrepreneuring," *Journal of Organizational Change Management*, vol. 11, no. 2, pp. 135–144, 1998.

[15] R. Isaak, "Globalisation and green entrepreneurship," *Greener Management International*, vol. 18, pp. 80–90, 1997.

[16] K. Hockerts and R. Wüstenhagen, "Greening Goliaths versus emerging Davids - Theorizing about the role of incumbents and new entrants in sustainable entrepreneurship," *Journal of Business Venturing*, vol. 25, no. 5, pp. 481–492, 2010.

[17] A. Triguero, L. Moreno-Mondéjar, and M. A. Davia, "Drivers of different types of eco-innovation in European SMEs," *Ecological Economics*, vol. 92, pp. 25–33, 2013.

[18] S.-P. Chuang and C.-L. Yang, "Key success factors when implementing a green-manufacturing system," *Production Planning and Control*, vol. 25, no. 11, pp. 923–937, 2014.

[19] J. G. Covin and G. T. Lumpkin, "Entrepreneurial orientation theory and research: reflections on a needed construct," *Entrepreneurship Theory and Practice*, vol. 35, no. 5, pp. 855–872, 2011.

[20] D. J. Teece, "Dynamic capabilities and entrepreneurial management in large organizations: toward a theory of the (entrepreneurial) firm," *European Economic Review*, vol. 86, pp. 202–216, 2016.

[21] K. Woldesenbet, M. Ram, and T. Jones, "Supplying large firms: The role of entrepreneurial and dynamic capabilities in small businesses," *International Small Business Journal*, vol. 30, no. 5, pp. 493–512, 2012.

[22] D. J. Teece, "A dynamic capabilities-based entrepreneurial theory of the multinational enterprise," *Journal of International Business Studies*, vol. 45, no. 1, pp. 8–37, 2014.

[23] Spanish Corporate Governance Code 1/2015 of 31 March. Retrieved: https://www.boe.es/boe/dias/2015/03/31/pdfs/BOE-A-2015-3439.pdf.

[24] F. Partnoy, *Infectious Greed: How Deceit and Risk Corrupted the Financial Markets*, Profile Books, London, UK, 2003.

[25] I. Melay and S. Kraus, "Green entrepreneurship: definitions of related concepts," *International Journal Strategegic Management*, vol. 12, pp. 1–12, 2012.

[26] A. V. Banerjee, E. Duflo, and K. Munshi, "The (MIS) allocation of capital," *Journal of the European Economic Association*, vol. 1, no. 2-3, pp. 484–494, 2003.

[27] W. L. Koe and I. A. Majid, "Socio-cultural factors and intention towards sustainable entrepreneurship," *Eurasian Journal of Business and Economics*, vol. 7, no. 13, pp. 145–156, 2014.

[28] D. Y. Choi and E. R. Gray, "The venture development processes of "sustainable" entrepreneurs," *Management Research News*, vol. 31, no. 8, pp. 558–569, 2008.

[29] Spanish Constitution, Congress of Deputies held on October 31. 1978. Retrieved. http://www.congreso.es/portal/page/portal/Congreso/Congreso/Hist_Normas/Norm/const_espa_texto_ingles_0.pdf.

[30] T. Gliedt and P. Parker, "Green community entrepreneurship: creative destruction in the social economy," *International Journal of Social Economics*, vol. 34, no. 8, pp. 538–553, 2007.

[31] I.-M. García-Sanchez, B. Cuadrado-Ballesteros, and J.-V. Frias-Aceituno, "Impact of the institutional macro context on the voluntary disclosure of CSR information," *Long Range Planning*, vol. 49, no. 1, pp. 15–35, 2016.

[32] J. M. Tamarit Sumalla, "La responsabilidad penal de las personas jurídicas," in *La reforma penal de 2010: análisis y comentarios, Quintero Olivares Dir., Cizur Menor*, p. 58, 2010.

[33] P. W. Moroz and K. Hindle, "Entrepreneurship as a process: toward harmonizing multiple perspectives," *Entrepreneurship Theory and Practice*, vol. 36, no. 4, pp. 781–818, 2012.

[34] J. A. Ingrisano and S. A. Mathews, "Practical guide to avoiding failure to supervise liability," *Preventive Law Reporter*, vol. 14, no. 12, 1995.

[35] K. B. Huff, "The role of corporate compliance programs in determining corporate criminal liability: a suggested approach," *Columbia Law Review*, vol. 96, no. 5, pp. 1252–1298, 1996.

[36] S. Sadiq, G. Governatori, and K. Namiri, "Modeling control objectives for business process compliance," in *Proceedings of the International Conference on Business Process Management*, pp. 149–164, Berlin, Germany, 2007.

[37] J. G. York and S. Venkataraman, "The entrepreneur-environment nexus: Uncertainty, innovation, and allocation," *Journal of Business Venturing*, vol. 25, no. 5, pp. 449–463, 2010.

[38] A. Amicelle, "Towards a new political anatomy of financial surveillance," *Security Dialogue*, vol. 42, no. 2, pp. 161–178, 2011.

[39] E. De Porres Ortiz de Urbina, *Responsabilidad Penal de las personas jurídicas*, El Derecho, Madrid, Spain, 2015.

[40] D. R. Campbell, M. Campbell, and G. W. Adams, "Adding significant value with internal controls," *The CPA Journal*, vol. 76, no. 6, p. 20, 2006.

[41] I. MacNeil and X. Li, "Comply or explain: market discipline and non-compliance with the combined code," *Corporate Governance: An International Review*, vol. 14, no. 5, pp. 486–496, 2006.

[42] R. Bampton and C. J. Cowton, "Taking stock of accounting ethics scholarship: a review of the journal literature," *Journal of Business Ethics*, vol. 114, no. 3, pp. 549–563, 2013.

[43] R. Calderón, I. Ferrero, and D. M. Redin, "Ethical codes and corporate responsibility of the most admired companies of the world: toward a third generation ethics?" *Business and Politics*, vol. 14, no. 4, pp. 1–24, 2012.

[44] M. P. Vandenbergh, "Beyond elegance: a testable typology of social norms in corporate environmental compliance," *Stanford Environmental Law Journal*, vol. 22, no. 55, 2003.

[45] B. V. Todeschini, M. N. Cortimiglia, D. Callegaro-de-Menezes, and A. Ghezzi, "Innovative and sustainable business models in the fashion industry: entrepreneurial drivers, opportunities, and challenges," *Business Horizons*, vol. 60, no. 6, pp. 759–770, 2017.

[46] K. Fletcher, "Slow fashion: an invitation for systems change," *The Journal of Design, Creative Process and the Fashion Industry*, vol. 2, no. 2, pp. 259–265, 2010.

[47] OCDE, "Sustainable development programmes and initiatives (2009-2010)," 2009. Retrieve. https://www.oecd.org/greengrowth/47445613.pdf.

[48] A. T. Bon and E. M. Mustafa, "Impact of total quality management on innovation in service organizations: literature review and new conceptual framework," *Procedia Engineering*, vol. 53, pp. 516–529, 2013.

[49] M. Mazzucato and C. C. R. Penna, *Beyond Market Failures: The Market Creating and Shaping Roles of State Investment Banks*, Levy Economics Institute, New York, NY, USA, 2015, (Working Paper October, 2015) Available at: http://www.levyinstitute.org/pubs/wp_831.pdf.

[50] J. Kirkwood and S. Walton, "How green is green? Ecopreneurs balancing environmental concerns and business goals," *Australasian Journal of Environmental Management*, vol. 21, no. 1, pp. 37–51, 2014.

[51] M. Shrivastava and J. P. Tamvada, "Which green matters for whom? Greening and firm performance across age and size distribution of firms," *Small Business Economics*, pp. 1–18, 2017.

[52] D. S. Dhaliwal, S. Radhakrishnan, A. Tsang, and Y. G. Yang, "Nonfinancial disclosure and analyst forecast accuracy: international evidence on corporate social responsibility disclosure," *The Accounting Review*, vol. 87, no. 3, pp. 723–759, 2012.

[53] P. M. Clarkson, Y. Li, G. D. Richardson, and F. P. Vasvari, "Revisiting the relation between environmental performance and environmental disclosure: an empirical analysis," *Accounting, Organizations and Society*, vol. 33, no. 4-5, pp. 303–327, 2008.

[54] D. F. Pacheco, T. J. Dean, and D. S. Payne, "Escaping the green prison: Entrepreneurship and the creation of opportunities for sustainable development," *Journal of Business Venturing*, vol. 25, no. 5, pp. 464–480, 2010.

[55] A. Pérez-Luño, J. Wiklund, and R. V. Cabrera, "The dual nature of innovative activity: How entrepreneurial orientation influences innovation generation and adoption," *Journal of Business Venturing*, vol. 26, no. 5, pp. 555–571, 2011.

[56] J. W. Carland, J. A. Carland, and J. W. Pearce, "Risk taking propensity among entrepreneurs, small business owners, and managers," *Journal of Business and Entrepreneurship*, vol. 7, no. 1, pp. 15–23, 1995.

[57] G. Shirokova, K. Bogatyreva, T. Beliaeva, and S. Puffer, "Entrepreneurial orientation and firm performance in different environmental settings: contingency and configurational approaches," *Journal of Small Business and Enterprise Development*, vol. 23, no. 3, pp. 703–727, 2016.

[58] M. Lenox and J. G. York, "Environmental entrepreneurship," in *Oxford Handbook of Business and the Environment*, A. J. Hoffman and T. Bansal, Eds., Oxford University Press, Oxford, UK, 2011.

[59] Pricewater House Coopers (PwC) Kuala Lumpur. Green tax incentives for Malaysia Oct; 86; 2010.

[60] D. J. Teece, "Dynamic capabilities: routines versus entrepreneurial action," *Journal of Management Studies*, vol. 49, no. 8, pp. 1395–1401, 2012.

[61] J. G. York, I. O'Neil, and S. D. Sarasvathy, "Exploring environmental entrepreneurship: identity coupling, venture goals, and stakeholder incentives," *Journal of Management Studies*, vol. 53, no. 5, pp. 695–737, 2016.

[62] M. Sarstedt, C. M. Ringle, and J. F. Hair, "Partial least squares structural equation modeling," in *Handbook of Market Research*, pp. 1–40, Springer International Publishing, 2017.

[63] C. B. Astrachan, V. K. Patel, and G. Wanzenried, "A comparative study of CB-SEM and PLS-SEM for theory development in family firm research," *Journal of Family Business Strategy*, vol. 5, no. 1, pp. 116–128, 2014.

[64] E. Carmines and R. Zeller, *Reliability and Validity Assessment*, vol. 17, SAGE Publications, Inc., Thousand Oaks, Calif, USA, 1979.

[65] O. Götz, K. Liehr-Gobbers, and M. Krafft, "Evaluation of structural equation models using the partial least squares (PLS) approach," in *Handbook of Partial Least Squares*, pp. 691–711, Springer, Berlin, Germany, 2010.

[66] J. Hair, W. Black, B. Babin, R. Anderson, and R. Tatham, *Multivariate data analysis*, Prentice Hall, Upper Saddle River, NJ, USA, 5th edition, 2005.

[67] C. Fornell and D. F. Larcker, "Structural equation models with unobservable variables and measurement error: algebra and statistics," *Journal of Marketing Research*, vol. 18, no. 3, pp. 382–388, 2018.

[68] T. K. Dijkstra and J. Henseler, "Consistent partial least squares path modeling," *MIS Quarterly: Management Information Systems*, vol. 39, no. 2, pp. 297–316, 2015.

[69] J. Henseler, C. M. Ringle, and R. R. Sinkovics, "The use of partial least squares path modeling in international marketing," in *New Challenges to International Marketing*, pp. 277–319, Emerald Group Publishing Limited, 2009.

[70] J. Henseler, C. M. Ringle, and M. Sarstedt, "A new criterion for assessing discriminant validity in variance-based structural equation modeling," *Journal of the Academy of Marketing Science*, vol. 43, no. 1, pp. 115–135, 2015.

[71] W. W. Chin, "The partial least squares approach to structural equation modeling," *Modern Methods for Business Research*, vol. 295, no. 2, pp. 295–336, 1998.

[72] L. T. Hu and P. M. Bentler, "Fit indices in covariance structure modeling: sensitivity to under-parameterized model misspecification," *Psychological Methods*, vol. 3, no. 4, pp. 424–453, 1998.

[73] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 36, pp. 111–147, 1974.

[74] S. Geisser, "A predictive approach to the random effect model," *Biometrika*, vol. 61, pp. 101–107, 1974.

[75] J.-M. Becker, A. Rai, C. M. Ringle, and F. Völckner, "Discovering unobserved heterogeneity in structural equation models to avert validity threats," *MIS Quarterly: Management Information Systems*, vol. 37, no. 3, pp. 665–694, 2013.

[76] M. Sarstedt, "A review of recent approaches for capturing heterogeneity in partial least squares path modelling," *Journal of Modelling in Management*, vol. 3, no. 2, pp. 140–161, 2008.

[77] P. Palos-Sanchez, F. Martin-Velicia, and J. R. Saura, "Complexity in the acceptance of sustainable search engines on the internet: an analysis of unobserved heterogeneity with FIMIX-PLS," *Complexity*, vol. 2018, Article ID 6561417, 19 pages, 2018.

[78] J. F. Hair, Jr., M. Sarstedt, L. M. Matthews, and C. M. Ringle, "Identifying and treating unobserved heterogeneity with FIMIX-PLS: part I – method," *European Business Review*, vol. 28, no. 1, pp. 63–76, 2016.

[79] J. Henseler, C. M. Ringle, and M. Sarstedt, "A new criterion for assessing discriminant validity in variance-based structural equation modeling," *Journal of the Academy of Marketing Science*, vol. 43, no. 1, pp. 115–135, 2014.

[80] M. Sarstedt, C. M. Ringle, D. Smith, R. Reams, and J. F. Hair, "Partial least squares structural equation modeling (PLS-SEM): a useful tool for family business researchers," *Journal of Family Business Strategy*, vol. 5, no. 1, pp. 105–115, 2014.

[81] M. Sarstedt, J. Becker, C. M. Ringle, and M. Schwaiger, "Uncovering and treating unobserved heterogeneity with FIMIX-PLS: which model selection criterion provides an appropriate number of segments?" *Schmalenbach Business Review*, vol. 63, no. 1, pp. 34–62, 2011.

[82] H. Bozdogan, "Mixture-model cluster analysis using model selection criteria and a new informational measure of complexity," in *Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach*, H. Bozdogan, Ed., pp. 69–113, Kluwer Academic Publishers, Boston, London, 1994.

[83] V. Ramaswamy, W. S. Desarbo, D. J. Reibstein, and W. T. Robinson, "An empirical pooling approach for estimating marketing mix elasticities with PIMS data," *Marketing Science*, vol. 12, no. 1, pp. 103–124, 1993.

[84] C. M. Ringle, S. Wende, and A. Will, "Customer segmentation with FIMIX-PLS," in *Proceedings of the PLS-05 International Symposium*, T. Aluja, J. Casanovas, and V. Esposito, Eds., pp. 507–514, PAD Test&Go, Paris, France, 2005.

[85] C. M. Ringle, S. Wende, and A. Will, "Finite mixture partial least squares analysis: methodology and numerical examples," in *Handbook of Partial Least Squares*, V. Esposito Vinzi, W. W. Chin, J. Henseler, and H. Wang, Eds., vol. 2 of *Springer handbooks of computational statistics series*, pp. 195–218, Springer, London, UK, 2010.

[86] L. Hu and P. M. Bentler, "Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives," *Structural Equation Modeling: A Multidisciplinary Journal*, vol. 6, no. 1, pp. 1–55, 1999.

[87] J. Mondéjar-Jiménez, M. Segarra-Oña, Á. Peiró-Signes, A. M. Payá-Martínez, and F. J. Sáez-Martínez, "Segmentation of the Spanish automotive industry with respect to the environmental orientation of firms: towards an ad-hoc vertical policy to promote eco-innovation," *Journal of Cleaner Production*, vol. 86, pp. 238–244, 2015.

[88] M. Tenenhaus, V. E. Vinzi, Y. Chatelin, and C. Lauro, "PLS path modeling," *Computational Statistics & Data Analysis*, vol. 48, no. 1, pp. 159–205, 2005.

[89] H. O. Wold, "Introduction to the second generation of multivariate analysis," in *Theoretical Empiricism: A General Rationale for Scientific Model-Building*, H. O. Wold, Ed., Paragon House, New York, NY, USA, 1989.

[90] W. Ching and P. Newsted, "Chapter 12. structural equation modeling. analysis with small samples using partial least squares," in *Statisticas Strategies for Smart Sample Researchs*, E. R. H. Hoyle, Ed., Sage Publications, Thousand Oaks, CA, USA, 1999.

[91] G. Shmueli and O. R. Koppius, "Predictive analytics in information systems research," *MIS Quarterly: Management Information Systems*, vol. 35, no. 3, pp. 553–572, 2011.

[92] C. M. Felipe, J. L. Roldán, and A. L. Leal-Rodríguez, "Impact of organizational culture values on organizational agility," *Sustainability*, vol. 9, no. 12, p. 2354, 2017.

[93] M. Sarstedt, C. M. Ringle, G. Schmueli, J. H. Cheah, and H. Ting, "Predictive model assessment in PLS-SEM: guidelines for using PLSpredict," *Working Paper*, 2018.

[94] C. M. Ringle, S. Wende, and J. M. Becker, "SmartPLS 3," Boenningstedt: SmartPLS GmbH, 2015. http://www.smartpls.com.

[95] N. Danks, S. Ray, and G. Shmueli, "Evaluating the predictive performance of constructs in PLS path modeling," *Working Paper*, 2018.

[96] A. G. Woodside, "Moving beyond multiple regression analysis to algorithms: Calling for adoption of a paradigm shift from symmetric to asymmetric thinking in data analysis and crafting theory," *Journal of Business Research*, vol. 66, no. 4, pp. 463–472, 2013.

[97] M. Groß, "Heterogeneity in consumers' mobile shopping acceptance: A finite mixture partial least squares modelling approach for exploring and characterising different shopper segments," *Journal of Retailing and Consumer Services*, vol. 40, pp. 8–18, 2018.

[98] A. Argandoña, "The united nations convention against corruption and its impact on international companies," *Journal of Business Ethics*, vol. 74, no. 4, pp. 481–496, 2007.

[99] J. Henseler, G. Hubona, and P. A. Ray, "Using PLS path modeling in new technology research: updated guidelines," *Industrial Management & Data Systems*, vol. 116, no. 1, pp. 2–20, 2016.

Hindawi
Complexity
Volume 2019, Article ID 4392785, 15 pages
https://doi.org/10.1155/2019/4392785

Research Article

A CEEMDAN and XGBOOST-Based Approach to Forecast Crude Oil Prices

Yingrui Zhou,[1] Taiyong Li,[1,2] Jiayi Shi,[1] and Zijie Qian[1]

[1]*School of Economic Information Engineering, Southwestern University of Finance and Economics, Chengdu 611130, China*
[2]*Sichuan Province Key Laboratory of Financial Intelligence and Financial Engineering, Southwestern University of Finance and Economics, Chengdu 611130, China*

Correspondence should be addressed to Taiyong Li; litaiyong@gmail.com

Received 13 October 2018; Accepted 13 January 2019; Published 3 February 2019

Guest Editor: Marisol B. Correia

Crude oil is one of the most important types of energy for the global economy, and hence it is very attractive to understand the movement of crude oil prices. However, the sequences of crude oil prices usually show some characteristics of nonstationarity and nonlinearity, making it very challenging for accurate forecasting crude oil prices. To cope with this issue, in this paper, we propose a novel approach that integrates complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) and extreme gradient boosting (XGBOOST), so-called CEEMDAN-XGBOOST, for forecasting crude oil prices. Firstly, we use CEEMDAN to decompose the nonstationary and nonlinear sequences of crude oil prices into several intrinsic mode functions (IMFs) and one residue. Secondly, XGBOOST is used to predict each IMF and the residue individually. Finally, the corresponding prediction results of each IMF and the residue are aggregated as the final forecasting results. To demonstrate the performance of the proposed approach, we conduct extensive experiments on the West Texas Intermediate (WTI) crude oil prices. The experimental results show that the proposed CEEMDAN-XGBOOST outperforms some state-of-the-art models in terms of several evaluation metrics.

1. Introduction

As one of the most important types of energy that power the global economy, crude oil has great impacts on every country, every enterprise, and even every person. Therefore, it is a crucial task for the governors, investors, and researchers to forecast the crude oil prices accurately. However, the existing research has shown that crude oil prices are affected by many factors, such as supply and demand, interest rate, exchange rate, speculation activities, international and political events, climate, and so on [1, 2]. Therefore, the movement of crude oil prices is irregular. For example, starting from about 11 USD/barrel in December 1998, the WTI crude oil prices gradually reached the peak of 145.31 USD/barrel in July 2008, and then the prices drastically declined to 30.28 USD/barrel in the next five months because of the subprime mortgage crisis. After that, the prices climbed to more than 113 USD/barrel in April 2011, and, once again, they sharply dropped to about 26 USD/barrel in February 2016. The movement of the crude oil prices in the last decades has shown that the forecasting task is very challenging, due to the characteristics of high nonlinearity and nonstationarity of crude oil prices.

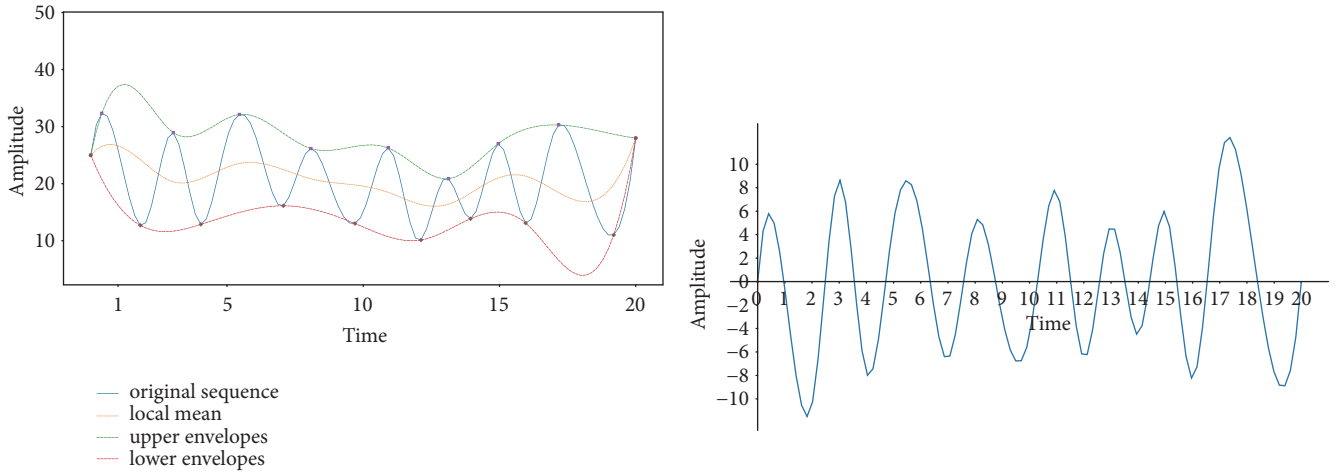
Many scholars have devoted efforts to trying to forecast crude oil prices accurately. The most widely used approaches to forecasting crude oil prices can be roughly divided into two groups: statistical approaches and artificial intelligence (AI) approaches. Recently, Miao et al. have explored the factors of affecting crude oil prices based on the least absolute shrinkage and selection operator (LASSO) model [1]. Ye et al. proposed an approach integrating ratchet effect for linear prediction of crude oil prices [3]. Morana put forward a semiparametric generalized autoregressive conditional heteroskedasticity (GARCH) model to predict crude oil prices at different lag periods even without the conditional average of historical crude oil prices [4]. Naser found that using the dynamic model averaging (DMA) with empirical evidence is better than linear models such as autoregressive (AR) model and its variants [5]. Gong and Lin proposed several new

heterogeneous autoregressive (HAR) models to forecast the good and bad uncertainties in crude oil prices [6]. Wen et al. also used HAR models with structural breaks to forecast the volatility of crude oil futures [7].

Although the statistical approaches improve the accuracy of forecasting crude oil prices to some extent, the assumption of linearity of crude oil prices cannot be met according to some recent research, and hence it limits the accuracy. Therefore, a variety of AI approaches have been proposed to capture the nonlinearity and nonstationarity of crude oil prices in the last decades [8–11]. Chiroma et al. reviewed the existing research associated with forecasting crude oil prices and found that AI methodologies are attracting unprecedented interest from scholars in the domain of crude oil price forecasting [8]. Wang et al. proposed an AI system framework that integrated artificial neural networks (ANN) and rule-based expert system with text mining to forecast crude oil prices, and it was shown that the proposed approach was significantly effective and practically feasible [9]. Barunik and Malinska used neural networks to forecast the term structure in crude oil futures prices [10]. Most recently, Chen et al. have studied forecasting crude oil prices using deep learning framework and have found that the random walk deep belief networks (RW-DBN) model outperforms the long short term memory (LSTM) and the random walk LSTM (RW-LSTM) models in terms of forecasting accuracy [11]. Other AI-methodologies, such as genetic algorithm [12], compressive sensing [13], least square support vector regression (LSSVR) [14], and cluster support vector machine (ClusterSVM) [15], were also applied to forecasting crude oil prices. Due to the extreme nonlinearity and nonstationarity, it is hard to achieve satisfactory results by forecasting the original time series directly. An ideal approach is to divide the tough task of forecasting original time series into several subtasks, and each of them forecasts a relatively simpler subsequence. And then the results of all subtasks are accumulated as the final result. Based on this idea, a "decomposition and ensemble" framework was proposed and widely applied to the analysis of time series, such as energy forecasting [16, 17], fault diagnosis [18–20], and biosignal analysis [21–23]. This framework consists of three stages. In the first stage, the original time series was decomposed into several components. Typical decomposition methodologies include wavelet decomposition (WD), independent component analysis (ICA) [24], variational mode decomposition (VMD) [25], empirical mode decomposition (EMD) [2, 26] and its extension (ensemble EMD (EEMD)) [27, 28], and complementary EEMD (CEEMD) [29]. In the second stage, some statistical or AI-based methodologies are applied to forecast each decomposed component individually. In theory, any regression methods can be used to forecast the results of each component. In the last stage, the predicted results from all components are aggregated as the final results. Recently, various researchers have devoted efforts to forecasting crude oil prices following the framework of "decomposition and ensemble." Fan et al. put forward a novel approach that integrates independent components analysis (ICA) and support vector regression (SVR) to forecast crude oil prices, and the experimental results verified the effectiveness of the proposed approach [24]. Yu et al. used EMD to decompose the sequences of the crude oil prices into several intrinsic mode functions (IMFs) at first and then used a three-layer feed-forward neural network (FNN) model for predicting each IMF. Finally, the authors used an adaptive linear neural network (ALNN) to combine all the results of the IMFS as the final forecasting output [2]. Yu et al. also used EEMD and extended extreme learning machine (EELM) to forecast crude oil prices, following the framework of "decomposition and ensemble." The empirical results demonstrated the effectiveness and efficiency of the proposed approach [28]. Tang et al. further proposed an improved approach integrating CEEMD and EELM for forecasting crude oil prices, and the experimental results demonstrated that the proposed approach outperformed all the listed state-of-the-art benchmarks [29]. Li et al. used EEMD to decompose raw crude oil prices into several components and then used kernel and nonkernel sparse Bayesian learning (SBL) to forecast each component, respectively [30, 31].

From the perspective of decomposition, although EMD and EEMD are capable of improving the accuracy of forecasting crude oil prices, they still suffer "mode mixing" and introducing new noise in the reconstructed signals, respectively. To overcome these shortcomings, an extension of EEMD, so-called complete EEMD with adaptive noise (CEEMDAN), was proposed by Torres et al. [32]. Later, the authors put forward an improved version of CEEMDAN to obtain decomposed components with less noise and more physical meaning [33]. The CEEMDAN has succeeded in wind speed forecasting [34], electricity load forecasting [35], and fault diagnosis [36–38]. Therefore, CEEMDAN may have the potential to forecast crude oil prices. As pointed out above, any regression methods can be used to forecast each decomposed component. A recently proposed machine learning algorithm, extreme gradient boosting (XGBOOST), can be used for both classification and regression [39]. The existing research has demonstrated the advantages of XGBOOST in forecasting time series [40–42].

With the potential of CEEMDAN in decomposition and XGBOOST in regression, in this paper, we aim at proposing a novel approach that integrates CEEMDAN and XGBOOST, namely, CEEMDAN-XGBOOST, to improve the accuracy of forecasting crude oil prices, following the "decomposition and ensemble" framework. Specifically, we firstly decompose the raw crude oil price series into several components with CEEMDAN. And then, for each component, XGBOOST is applied to building a specific model to forecast the component. Finally, all the predicted results from every component are aggregated as the final forecasting results. The main contributions of this paper are threefold: (1) We propose a novel approach, so-called CEEMDAN-XGBOOST, for forecasting crude oil prices, following the "decomposition and ensemble" framework; (2) extensive experiments are conducted on the publicly-accessed West Texas Intermediate (WTI) to demonstrate the effectiveness of the proposed approach in terms of several evaluation metrics; (3) we further study the impacts of several parameter settings with the proposed approach.



(a) The line of the original sequence, local mean, upper envelopes, and lower envelopes

(b) The first IMF decomposed by EMD from the above original sequence

Figure 1: An illustration of EMD.

The remainder of this paper is organized as follows. Section 2 describes CEEMDAN and XGBOOST. Section 3 formulates the proposed CEEMDAN-XGBOOST approach in detail. Experimental results are reported and analyzed in Section 4. We also discussed the impacts of parameter settings in this section. Finally, Section 5 concludes this paper.

2. Preliminaries

2.1. EMD, EEMD and CEEMDAN. EMD was proposed by Huang et al. in 1998, and it has been developed and applied in many disciplines of science and engineering [26]. The key feature of EMD is to decompose a nonlinear, nonstationary sequence into intrinsic mode functions (IMFs) in the spirit of the Fourier series. In contrast to the Fourier series, they are not simply sine or cosine functions, but rather functions that represent the characteristics of the local oscillation frequency of the original data. These IMFs need to satisfy two conditions: (1) the number of local extrema and the number of zero crossing must be equal or differ at most by one and (2) the curve of the "local mean" is defined as zero.

At first, EMD finds out the upper and the lower envelopes which are computed by finding the local extrema of the original sequence. Then, the local maxima (minima) are linked by two cubic spines to construct the upper (lower) envelopes, respectively. The mean of these envelopes is considered as the "local mean." Meanwhile, the curve of this "local mean" is defined as the first residue, and the difference between original sequence and the "local mean" is defined as the first IMF. An illustration of EMD is shown in Figure 1.

After the first IMF is decomposed by EMD, there is still a residue (the local mean, i.e., the yellow dot line in Figure 1(a)) between the IMF and the original data. Obviously, extrema and high-frequency oscillations also exist in the residue. And EMD decomposes the residue into another IMF and one residue. If the variance of the new residue is not small enough to satisfy the Cauchy criterion, EMD will repeat to decompose new residue into another IMF and a new residue. Finally, EMD decomposes original sequence into several IMFs and one residue. The difference between the IMF and the residues is defined as

$$r_k[t] = r_{k-1}[t] - IMF_k[t], \quad k = 2, \ldots, K, \tag{1}$$

where $r_k[t]$ is the k-th residue at the time t and K is the total number of IMFs and residues.

Subsequently, Huang et al. thought that EMD could not extract the local features from the mixed features of the original sequence completely. One of the reasons for this is the frequent appearance of the mode mixing. The mode mixing can be defined as the situation that similar pieces of oscillations exist at the same corresponding position in different IMFs, which causes that a single IMF has lost its physical meanings. What is more, if one IMF has this problem, the following IMFs cannot avoid it either. To solve this problem, Wu and Huang extended EMD to a new version, namely, EEMD, that adds white noise to the original time series and performs EMD many times [27]. Given a time series and corresponding noise, the new time series can be expressed as

$$x^i[t] = x[t] + w^i[t], \tag{2}$$

where $x[t]$ stands for the original data and $w^i[t]$ is the i-th white noise (i=1,2,...,N, and N is the times of performing EMD).

Then, EEMD decomposes every $x^i[t]$ into $IMF_k^i[t]$. In order to get the real k-th IMF, $\overline{IMF_k}$, EEMD calculates the average of the $IMF_k^i[t]$. In theory, because the mean of the white noise is zero, the effect of the white noise would be eliminated by computing the average of $IMF_k^i[t]$, as shown in

$$\overline{IMF_k} = \frac{1}{N}\sum_{i=1}^{N} IMF_k^i[t]. \tag{3}$$

However, Torres et al. found that, due to the limited number of $x^i[t]$ in empirical research, EEMD could not completely eliminate the influence of white noise in the end. For this situation, Torres et al. put forward a new decomposition technology, CEEMDAN, on the basis of EEMD [32].

CEEMDAN decomposes the original sequence into the first IMF and residue, which is the same as EMD. Then, CEEMDAN gets the second IMF and residue, as shown in

$$\overline{IMF_2} = \frac{1}{N}\sum_{i=1}^{N} E_1\left(r_1[t] + \varepsilon_1 E_1\left(w^i[t]\right)\right), \tag{4}$$

$$r_2[t] = r_1[t] - \overline{IMF_2}, \tag{5}$$

where $E_1(.)$ stands for the first IMF decomposed from the sequence and ε_i is used to set the signal-to-noise ratio (SNR) at each stage.

In the same way, the k-th IMF and residue can be calculated as

$$\overline{IMF_k} = \frac{1}{N}\sum_{i=1}^{N} E_1\left(r_{k-1}[t] + \varepsilon_{k-1} E_{k-1}\left(w^i[t]\right)\right), \tag{6}$$

$$r_k[t] = r_{k-1}[t] - \overline{IMF_k}, \tag{7}$$

Finally, CEEMDAN gets several IMFs and computes the residue, as shown in

$$\mathrm{R}[t] = \mathrm{x}[t] - \sum_{k=1}^{K} \overline{IMF_k}. \tag{8}$$

The sequences decomposed by EMD, EEMD, and CEEMDAN satisfy (8). Although CEEMDAN can solve the problems that EEMD leaves, it still has two limitations: (1) the residual noise that the models contain and (2) the existence of spurious modes. Aiming at dealing with these issues, Torres et al. proposed a new algorithm to improve CEEMDAN [33].

Compared with the original CEEMDAN, the improved version obtains the residues by calculating the local means. For example, in order to get the first residue shown in (9), it would compute the local means of N realizations $x^i[t] = x[t] + \varepsilon_0 E_1(w^i[t])(i=1, 2,..., N)$.

$$r_1[t] = \frac{1}{N}\sum_{i=1}^{N} M\left(x^i[t]\right), \tag{9}$$

where $M(.)$ is the local mean of the sequence.

Then, it can get the first IMF shown in

$$IMF_1 = x[t] - r_1[t]. \tag{10}$$

For the k-th residue and IMF, they can be computed as (11) and (12), respectively:

$$r_k[t] = \frac{1}{N}\sum_{i=1}^{N} M\left(r_{k-1}[t] + \varepsilon_{k-1} E_{\mathrm{k}}\left(w^i[t]\right)\right), \tag{11}$$

$$IMF_k = r_{k-1}[t] - r_k[t]. \tag{12}$$

The authors have demonstrated that the improved CEEMDAN outperformed the original CEEMDAN in signal decomposition [33]. In what follows, we will refer to the improved version of CEEMDAN as CEEMDAN, unless otherwise stated. With CEEMDAN, the original sequence can be decomposed into several IMFs and one residue, that is, the tough task of forecasting the raw time series, can be divided into forecasting several simpler subtasks.

2.2. XGBOOST. Boosting is the ensemble method that can combine several weak learners into a strong learner as

$$\widehat{y_i} = \emptyset(x_i) = \sum_{k=1}^{K} f_k(x_i), \tag{13}$$

where $f_k(.)$ is a weak learner and K is the number of weak learners.

When it comes to the tree boosting, its learners are decision trees which can be used for both regression and classification.

To a certain degree, XGBOOST is considered as tree boost, and its core is the Newton boosting instead of Gradient Boosting, which finds the optimal parameters by minimizing the loss function (∅), as shown in

$$L(\emptyset) = \sum_{i=1}^{n} l\left(\widehat{y_i}, y_i\right) + \sum_{k=1}^{K} \Omega\left(f_k\right), \tag{14}$$

$$\Omega\left(f_k\right) = \gamma T + \frac{1}{2}\alpha\|\omega\|^2, \tag{15}$$

where $\Omega(f_k)$ is the complexity of the k-th tree model, n is the sample size, T is the number of leaf nodes of the decision trees, ω is the weight of the leaf nodes, γ controls the extent of complexity penalty for tree structure on T, and α controls the degree of the regularization of f_k.

Since it is difficult for the tree ensemble model to minimize loss function in (14) and (15) with traditional methods in Euclidean space, the model uses the additive manner [43]. It adds f_t that improves the model and forms the new loss function as

$$L^t = \sum_{i=1}^{n} l\left(y_i, \widehat{y_i}^{(t-1)} + f_t(x_i)\right) + \Omega\left(f_t\right), \tag{16}$$

where $\widehat{y_i}^{(t)}$ is the prediction of the i-th instance at the t-th iteration and f_t is the weaker learner at the t-th iteration.

Then, Newton boosting performs a Taylor second-order expansion on the loss function $l(y_i, \widehat{y_i}^{(t-1)} + f_t(x_i))$ to obtain $g_i f_t(x_i) + (1/2)h_i f_t^2(x_i)$, because the second-order approximation helps to minimize the loss function conveniently and quickly [43]. The equations of g_i, h_i and the new loss function are defined, respectively, as

$$g_i = \frac{\partial l\left(y_i, \widehat{y_i}^{(t-1)}\right)}{\widehat{y_i}^{(t-1)}}, \tag{17}$$

$$h_i = \frac{\partial^2 l\left(y_i, \widehat{y}_i^{(t-1)}\right)}{\widehat{y}_i^{(t-1)}}, \tag{18}$$

$$L_t = \sum_{i=1}^{n}\left[l\left(y_i, \widehat{y}_i^{(t-1)}\right) + g_i f_t\left(x_i\right) + \frac{1}{2}h_i f_t^2\left(x_i\right)\right] + \Omega\left(f_t\right). \tag{19}$$

Assume that the sample set I_j in the leaf node j is defined as $I_j=\{i \mid q(x_i) = j\}$, where $\mathrm{q}(x_i)$ represents the tree structure from the root to the leaf node j in the decision tree, (19) can be transformed into the following formula, as show in

$$\begin{aligned}\widetilde{L}^{(t)} &= \sum_{i=1}^{n}\left[g_i f_t\left(x_i\right) + \frac{1}{2}h_i f_t^2\left(x_i\right)\right] + \gamma T + \frac{1}{2}\alpha\sum_{j=1}^{T}\omega_j^2 \\ &= \sum_{j=1}^{T}\left[\left(\sum_{i\in I_j} g_i\right)\omega_j + \frac{1}{2}\left(\sum_{i\in I_j} h_i + a\right)\omega_j^2\right] + \gamma T.\end{aligned} \tag{20}$$

The formula for estimating the weight of each leaf in the decision tree is formulated as

$$\omega_j^* = -\frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i + a}, \tag{21}$$

According to (21), as for the tree structure q, the loss function at the leaf node j can be changed as

$$\widetilde{L}^{(t)}\left(q\right) = -\frac{1}{2}\sum_{j=1}^{T}\frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j} h_i + a} + \gamma T, \tag{22}$$

Therefore, the equation of the information gain after branching can be defined as

$$Gain = \frac{1}{2}\left[\frac{\left(\sum_{i\in I_{jL}} g_i\right)^2}{\sum_{i\in I_{jL}} h_i + a} + \frac{\left(\sum_{i\in I_{jR}} g_i\right)^2}{\sum_{i\in I_{jR}} h_i + a} - \frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j} h_i + a}\right] - \gamma, \tag{23}$$

where I_{jL} and I_{jR} are the sample sets of the left and right leaf node, respectively, after splitting the leaf node j.

XGBOOST branches each leaf node and constructs basic learners by the criterion of maximizing the information gain.

With the help of Newton boosting, the XGBOOST can deal with missing values by adaptively learning. To a certain extent, XGBOOST is based on the multiple additive regression tree (MART), but it can get better tree structure by learning with Newton boosting. In addition, XGBOOST can also subsample among columns, which reduces the relevance of each weak learner [39].

3. The Proposed CEEMDAN-XGBOOST Approach

From the existing literature, we can see that CEEMDAN has advantages in time series decomposition, while XGBOOST does well in regression. Therefore, in this paper, we integrated these two methods and proposed a novel approach, so-called CEEMDAN-XGBOOST, for forecasting crude oil prices. The proposed CEEMDAN-XGBOOST includes three stages: decomposition, individual forecasting, and ensemble. In the first stage, CEEMDAN is used to decompose the raw series of crude oil prices into k+1 components, including k IMFs and one residue. Among the components, some show high-frequency characteristics while the others show low-frequency ones of the raw series. In the second stage, for each component, a forecasting model is built using XGBOOST, and then the built model is applied to forecast each component and then get an individual result. Finally, all the results from the components are aggregated as the final result. Although there exist a lot of methods to aggregate the forecasting results from components, in the proposed approach, we use the simplest way, i.e., addition, to summarize the results of all components. The flowchart of the CEEMDAN-XGBOOST is shown in Figure 2.

From Figure 2, it can be seen that the proposed CEEMDAN-XGBOOST based on the framework of "decomposition and ensemble" is also a typical strategy of "divide and conquer"; that is, the tough task of forecasting crude oil prices from the raw series is divided into several subtasks of forecasting from simpler components. Since the raw series is extremely nonlinear and nonstationary while each decomposed component has a relatively simple form for forecasting, the CEEMDAN-XGBOOST has the ability to achieve higher accuracy of forecasting crude oil prices. In short, the advantages of the proposed CEEMDAN-XGBOOST are threefold: (1) the challenging task of forecasting crude oil prices is decomposed into several relatively simple subtasks; (2) for forecasting each component, XGBOOST can build models with different parameters according to the characteristics of the component; and (3) a simple operation, addition, is used to aggregate the results from subtasks as the final result.

4. Experiments and Analysis

4.1. Data Description. To demonstrate the performance of the proposed CEEMDAN-XGBOOST, we use the crude oil prices from the West Texas Intermediate (WTI) as experimental data (the data can be downloaded from https://www.eia.gov/dnav/pet/hist/RWTCD.htm). We use the daily closing prices covering the period from January 2, 1986, to March 19, 2018, with 8123 observations in total for empirical studies. Among the observations, the first 6498 ones from January 2, 1986, to September 21, 2011, accounting for 80% of the total observations, are used as training samples, while the remaining 20% ones are for testing. The original crude oil prices are shown in Figure 3.

We perform multi-step-ahead forecasting in this paper. For a given time series x_t $(t = 1, 2, \ldots, T)$, the m-step-ahead forecasting for x_{t+m} can be formulated as

$$\widehat{x}_{t+m} = f\left(x_{t-(l-1)}, x_{t-(l-2)}, \ldots, x_{t-1}, x_t\right), \tag{24}$$

where $\widehat{x}_{t+m}$ is the m-step-ahead predicted result at time t, f is the forecasting model, x_i is the true value at time i, and l is the lag order.

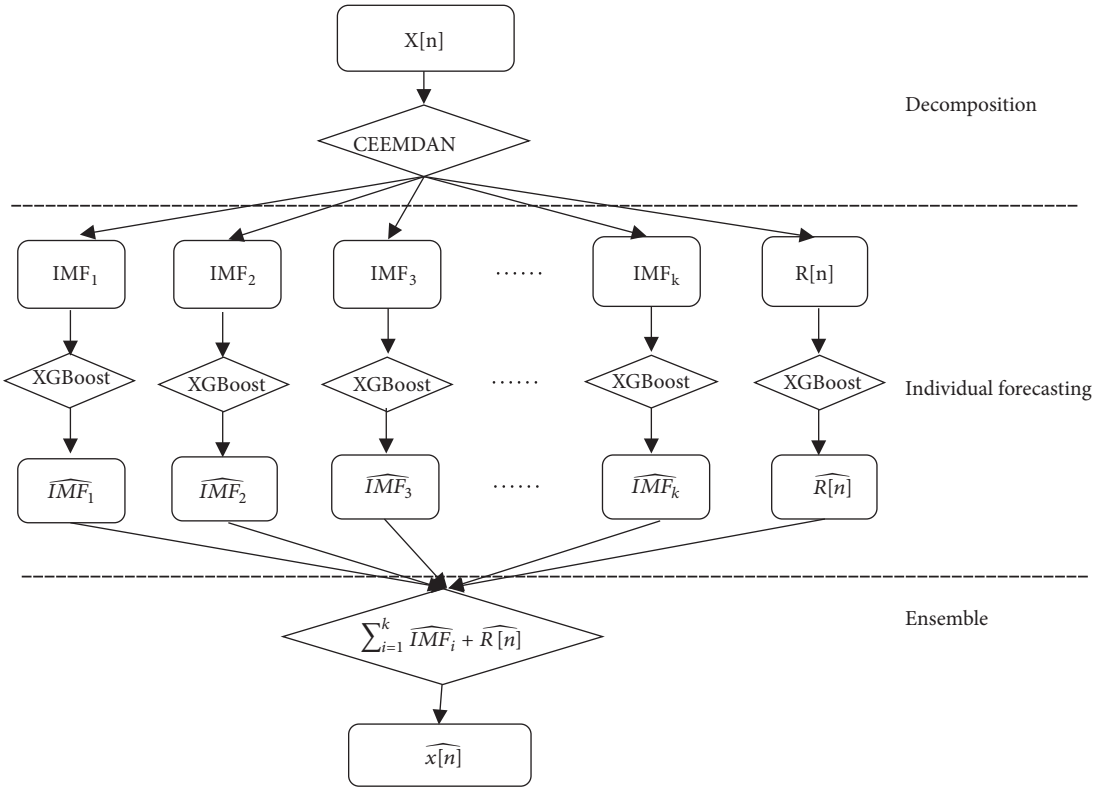


Figure 2: The flowchart of the proposed CEEMDAN-XGBOOST.

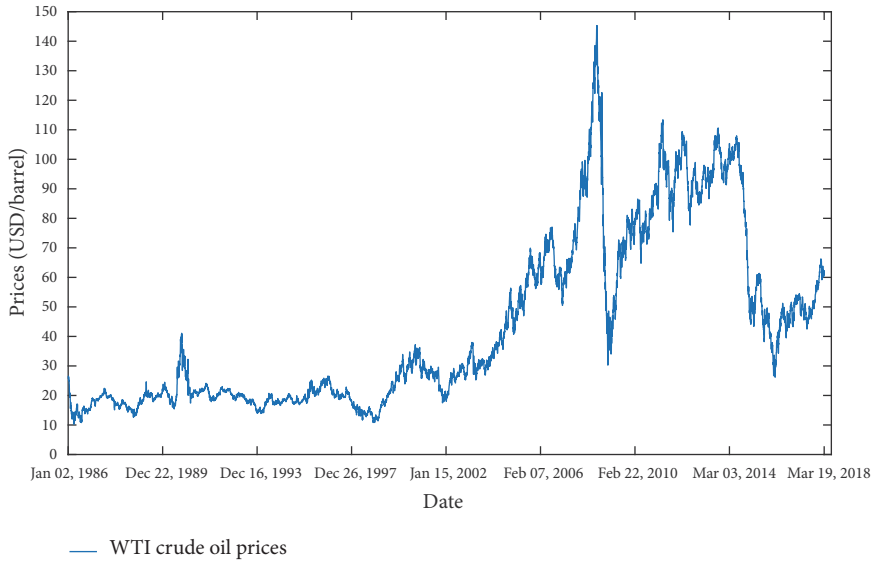


Figure 3: The original crude oil prices of WTI.

For SVR and FNN, we normalize each decomposed component before building the model to forecast the component individually. In detail, the normalization process can be defined as

$$x'_t = \frac{x_t - \mu}{\sigma}, \tag{25}$$

where x'_t is the normalized series of crude oil prices series, x_t is the data before normalization, μ is the mean of x_t, and σ is the standard deviation of x_t. Meanwhile, since normalization is not necessary for XGBOOST and ARIMA, for models with these two algorithms, we build forecasting models from each of the decomposed components directly.

4.2. Evaluation Criteria. When we evaluate the accuracy of the models, we focus on not only the numerical accuracy but also the accuracy of forecasting direction. Therefore, we select the root-mean-square error (RMSE) and the mean absolute error (MAE) to evaluate the numerical accuracy of the models. Besides, we use directional statistic (Dstat) as the criterion for evaluating the accuracy of forecasting direction. The RMSE, MAE, and Dstat are defined as

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_t - \widehat{y}_t\right)^2}, \tag{26}$$

$$\text{MAE} = \frac{1}{N}\left(\sum_{i=1}^{N}\left|y_t - \widehat{y}_t\right|\right), \tag{27}$$

$$\text{Dstat} = \frac{\sum_{i=2}^{N} d_i}{N-1}, \quad d_i = \begin{cases} 1, & \left(y_t - y_{t-1}\right)\left(\widehat{y}_t - \widehat{y_{t-1}}\right) > 0 \\ 0, & \left(y_t - y_{t-1}\right)\left(\widehat{y}_t - \widehat{y_{t-1}}\right) < 0, \end{cases} \tag{28}$$

where y_t is the actual crude oil prices at the time t,$\widehat{y}_t$ is the prediction, and N is the size of the test set.

In addition, we take the Wilcoxon signed rank test (WSRT) for proving the significant differences between the forecasts of the selected models [44]. The WSRT is a nonparametric statistical hypothesis test that can be used to evaluate whether the population mean ranks of two predictions from different models on the same sample differ. Meanwhile, it is a paired difference test which can be used as an alternative to the paired Student t-test. The null hypothesis of the WSRT is whether the median of the loss differential series $\mathrm{d(t)} = \mathrm{g}(e_a(\mathrm{t})) - \mathrm{g}(e_b(\mathrm{t}))$ is equal to zero or not, where e_a(t) and e_b(t) are the error series of model a and model b respectively, and g(.) is a loss function. If the p value of pairs of models is below 0.05, the test rejects the null hypothesis (there is a significant difference between the forecasts of this pair of models) under the confidence level of 95%. In this way, we can prove that there is a significant difference between the optimal model and the others.

However, the criteria defined above are global. If some singular points exist, the optimal model chosen by these criteria may not be the best one. Thus, we make the model confidence set (MCS) [31, 45] in order to choose the optimal model convincingly.

In order to calculate the p-value of the statistics accurately, the MCS performs bootstrap on the prediction series, which can soften the impact of the singular points. For the j-th model, suppose that the size of a bootstrapped sample is T, and the t-th bootstrapped sample has the loss functions defined as

$$L_{j,t} = \frac{1}{T}\sum_{t=h+1}^{h+T}\left\lfloor y_t - \widehat{y}_t \right\rfloor, \tag{29}$$

Suppose that a set $\mathrm{M}_0=\{\mathrm{m_i},\ \mathrm{i} = 1, 2, 3, \ldots, \mathrm{n}\}$ that contains n models, for any two models j and k, the relative values of the loss between these two models can be defined as

$$d_{j,k,t} = L_{j,t} - L_{k,t}, \tag{30}$$

According to the above definitions, the set of superior models can be defined as

$$M^* \equiv \left\{m_j \in \mathrm{M}_0 : E\left(d_{j,k,t}\right) \le 0,\ \forall m_k \in \mathrm{M}_0\right\}, \tag{31}$$

where E(.) represents the average value.

The MCS repeatedly performs the significant test in M_0. At each time, the worst prediction model in the set is eliminated. In the test, the hypothesis is the null hypothesis of equal predictive ability (EPA), defined as

$$H_0 : E\left(d_{j,k,t}\right) = 0, \quad \forall m_j, m_k \in M \subset \mathrm{M}_0 \tag{32}$$

The MCS mainly depends on the equivalence test and elimination criteria. The specific process is as follows.

Step 1. Assuming $\mathrm{M}=M_0$, at the level of significance α, use the equivalence test to test the null hypothesis $H_{0,M}$.

Step 2. If it accepts the null hypothesis and then it defines $M^*_{1-\alpha} = M$, otherwise it eliminates the model which rejects the null hypothesis from M according to the elimination criteria. The elimination will not stop until there are not any models that reject the null hypothesis in the set M. In the end, the models in $M^*_{1-\alpha}$ are thought as surviving models.

Meanwhile, the MCS has two kinds of statistics that can be defined as

$$T_R = \max_{j,k\in M}\left|t_{j,k}\right|, \tag{33}$$

$$T_{SQ} = \max_{j,k\in \mathrm{M}} t^2_{j,k}, \tag{34}$$

$$t_{j,k} = \frac{\overline{d_{j,k}}}{\sqrt{\mathrm{var}\left(\overline{d_{j,k}}\right)}}, \tag{35}$$

$$\overline{d_{j,k}} = \frac{1}{T}\sum_{t=h+1}^{h+T} d_{j,k,t}, \tag{36}$$

where T_R and T_{SQ} stand for the range statistics and the semiquadratic statistic, respectively, and both statistics are based on the t-statistics as shown in (35)-(36). These two statistics (T_R and T_{SQ}) are mainly to remove the model whose p-value is less than the significance level α. When the p-value is greater than the significance level α, the models can survive. The larger the p-value, the more accurate the prediction of the model. When the p-value is equal to 1, it indicates that the model is the optimal forecasting model.

4.3. Parameter Settings. To test the performance of XGBOOST and CEEMDAN-XGBOOST, we conduct two groups of experiments: single models that forecast crude

Table 1: The ranges of the parameters for XGBOOST by grid search.

Parameter	Description	range
Booster	Booster to use.	{'gblinear', 'gbtree'}
N_estimators	Number of boosted trees.	{100,200,300,400,500}
Max_depth	Maximum tree depth for base learners.	{3,4,5,6,7,8}
Min_child_weight	Maximum delta step we allow each tree's weight estimation to be.	{1,2,3,4,5,6}
Gamma	Minimum loss reduction required to make a further partition on a leaf node of the tree.	{0.01, 0.05,0.1,0.2,0.3 }
Subsample	Subsample ratio of the training instance.	{0.6,0.7,0.8,0.9,1}
Colsample	Subsample ratio of columns when constructing each tree.	{0.6,0.7,0.8,0.9,1}
Reg_alpha	L1 regularization term on weights	{0.01,0.05,0.1}
Reg_lambda	L2 regularization term on weights	{0.01,0.05,0.1}
Learning_rate	Boosting learning rate	{0.01,0.05,0.07,0.1,0.2}

oil prices with original sequence, and ensemble models that forecast crude oil prices based on the "decomposition and ensemble" framework.

For single models, we compare XGBOOST with one statistical model, ARIMA, and two widely used AI-models, SVR and FNN. Since the existing research has shown that EEMD significantly outperforms EMD in forecasting crude oil prices [24, 31], in the experiments, we only compare CEEMDAN with EEMD. Therefore, we compare the proposed CEEMDAN-XGBOOST with EEMD-SVR, EEMD-FNN, EEMD-XGBOOST, CEEMDAN-SVR, and CEEMDAN-FNN.

For ARIMA, we use the Akaike information criterion (AIC) [46] to select the parameters (p-d-q). For SVR, we use RBF as kernel function and use grid search to optimize C and *gamma* in the ranges of $2^{\{0,1,2,3,4,5,6,7,8\}}$ and $2^{\{-9,-8,-7,-6,-5,-4,-3,-2,-1,0\}}$, respectively. We use one hidden layer with 20 nodes for FNN. We use a grid search to optimize the parameters for XGBOOST; the search ranges of the optimized parameters are shown in Table 1.

We set 0.02 and 0.05 as the standard deviation of the added white noise and set 250 and 500 as the number of realizations of EEMD and CEEMDAN, respectively. The decomposition results of the original crude oil prices by EEMD and CEEMDAN are shown in Figures 4 and 5, respectively.

It can be seen from Figure 4 that, among the components decomposed by EEMD, the first six IMFs show characteristics of high frequency while the remaining six components show characteristics of low frequency. However, regarding the components by CEEMDAN, the first seven ones show clear high-frequency and the last four show low-frequency, as shown in Figure 5.

The experiments were conducted with Python 2.7 and MATLAB 8.6 on a 64-bit Windows 7 with 3.4 GHz I7 CPU and 32 GB memory. Specifically, we run FNN and MCS with MATLAB, and, for the remaining work, we use Python. Regarding XGBoost, we used a widely used Python package (https://xgboost.readthedocs.io/en/latest/python/) in the experiments.

Table 2: The RMSE, MAE, and Dstat by single models with horizon = 1, 3, and 6.

Horizon	Model	RMSE	MAE	Dstat
1	XGBOOST	**1.2640**	**0.9481**	0.4827
	SVR	1.2899	0.9651	0.4826
	FNN	1.3439	0.9994	0.4837
	ARIMA	1.2692	0.9520	**0.4883**
3	XGBOOST	**2.0963**	**1.6159**	0.4839
	SVR	2.2444	1.7258	**0.5080**
	FNN	2.1503	1.6512	0.4837
	ARIMA	2.1056	1.6177	0.4901
6	XGBOOST	**2.9269**	2.2945	0.5158
	SVR	3.1048	2.4308	**0.5183**
	FNN	3.0803	2.4008	0.5028
	ARIMA	2.9320	**2.2912**	0.5151

4.4. Experimental Results. In this subsection, we use a fixed value 6 as the lag order, and we forecast crude oil prices with 1-step-ahead, 3-step-ahead, and 6-step-ahead forecasting; that is to say, the horizons for these three forecasting tasks are 1, 3, and 6, respectively.

4.4.1. Experimental Results of Single Models. For single models, we compare XGBOOST with state-of-the-art SVR, FNN, and ARIMA, and the results are shown in Table 2.

It can be seen from Table 2 that XGBOOST outperforms other models in terms of RMSE and MAE with horizons 1 and 3. For horizon 6, XGBOOST achieves the best RMSE and the second best MAE, which is slightly worse than that by ARIMA. For horizon 1, FNN achieves the worst results among the four models; however, for horizons 3 and 6, SVR achieves the worst results. Regarding Dstat, none of the models can always outperform others, and the best result of Dstat is achieved by SVR with horizon 6. It can be found that the RMSE and MAE values gradually increase with the increase of horizon. However, the Dstat values do not show such discipline. All the values of Dstat are around 0.5, i.e., from

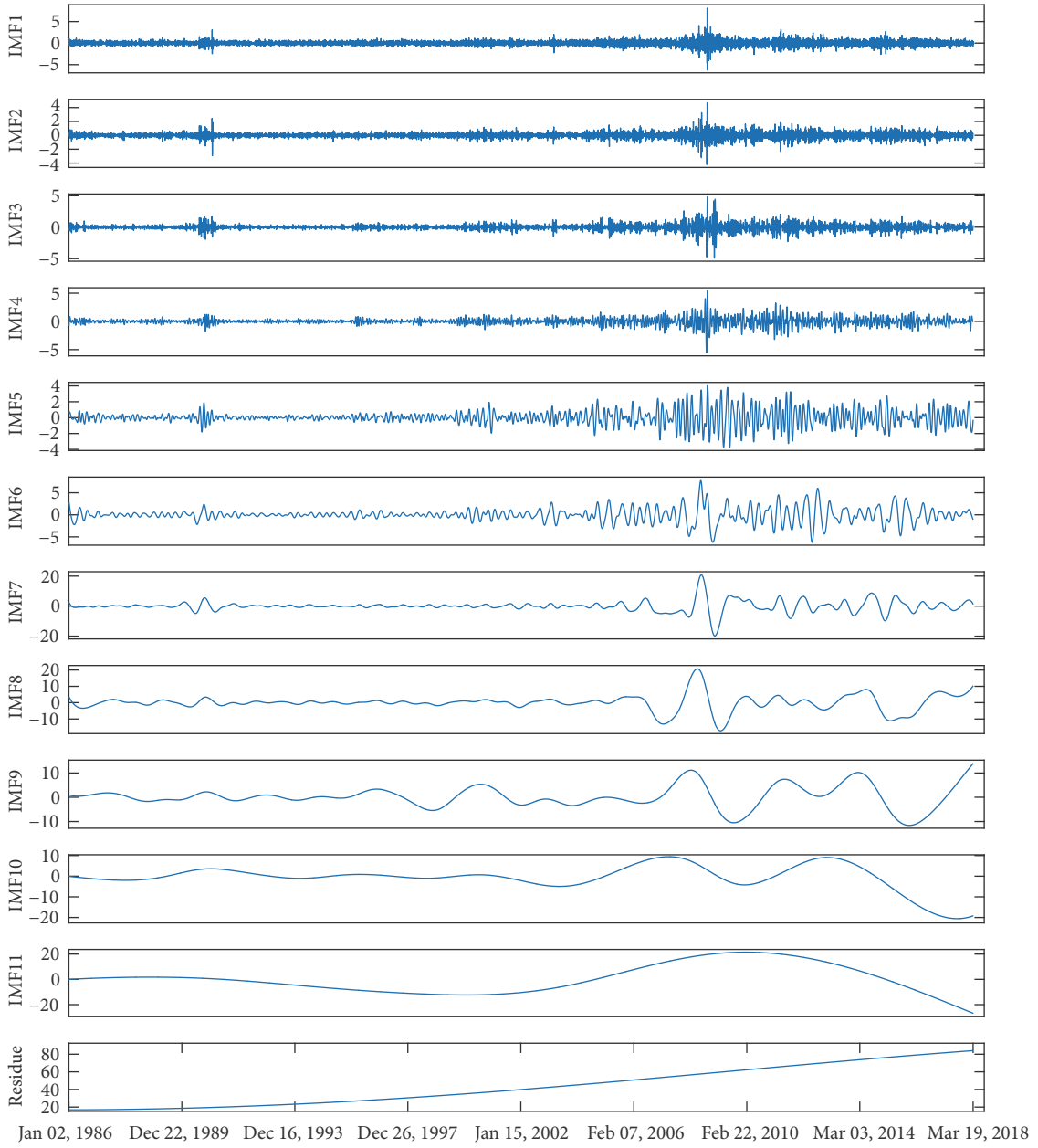


Figure 4: The IMFs and residue of WTI crude oil prices by EEMD.

0.4826 to 0.5183, which are very similar to the results of random guesses, showing that it is very difficult to accurately forecast the direction of the movement of crude oil prices with raw crude oil prices directly.

To further verify the advantages of XGBOOST over other models, we report the results by WSRT and MCS, as shown in Tables 3 and 4, respectively. As for WSRT, the p-value between XGBOOST and other models except ARIMA is below 0.05, which means that there is a significant difference among the forecasting results of XGBOOST, SVR, and FNN in population mean ranks. Besides, the results of MCS show that the p-value of T_R and T_{SQ} of XGBOOST is always equal to 1.000 and prove that XGBOOST is the optimal model among all the models in terms of global errors and most local errors of different samples obtained by bootstrap methods in MCS. According to MCS, the p-values of T_R and T_{SQ} of SVR on the horizon of 3 are greater than 0.2, so SVR becomes the

Table 3: Wilcoxon signed rank test between XGBOOST, SVR, FNN, and ARIMA.

	XGBOOST	SVR	FNN	ARIMA
XGBOOST	1	4.0378e-06	2.2539e-35	5.7146e-01
SVR	4.0378e-06	1	4.6786e-33	0.7006
FNN	2.2539e-35	4.6786e-33	1	6.9095e-02
ARIMA	5.7146e-01	0.7006	6.9095e-02	1

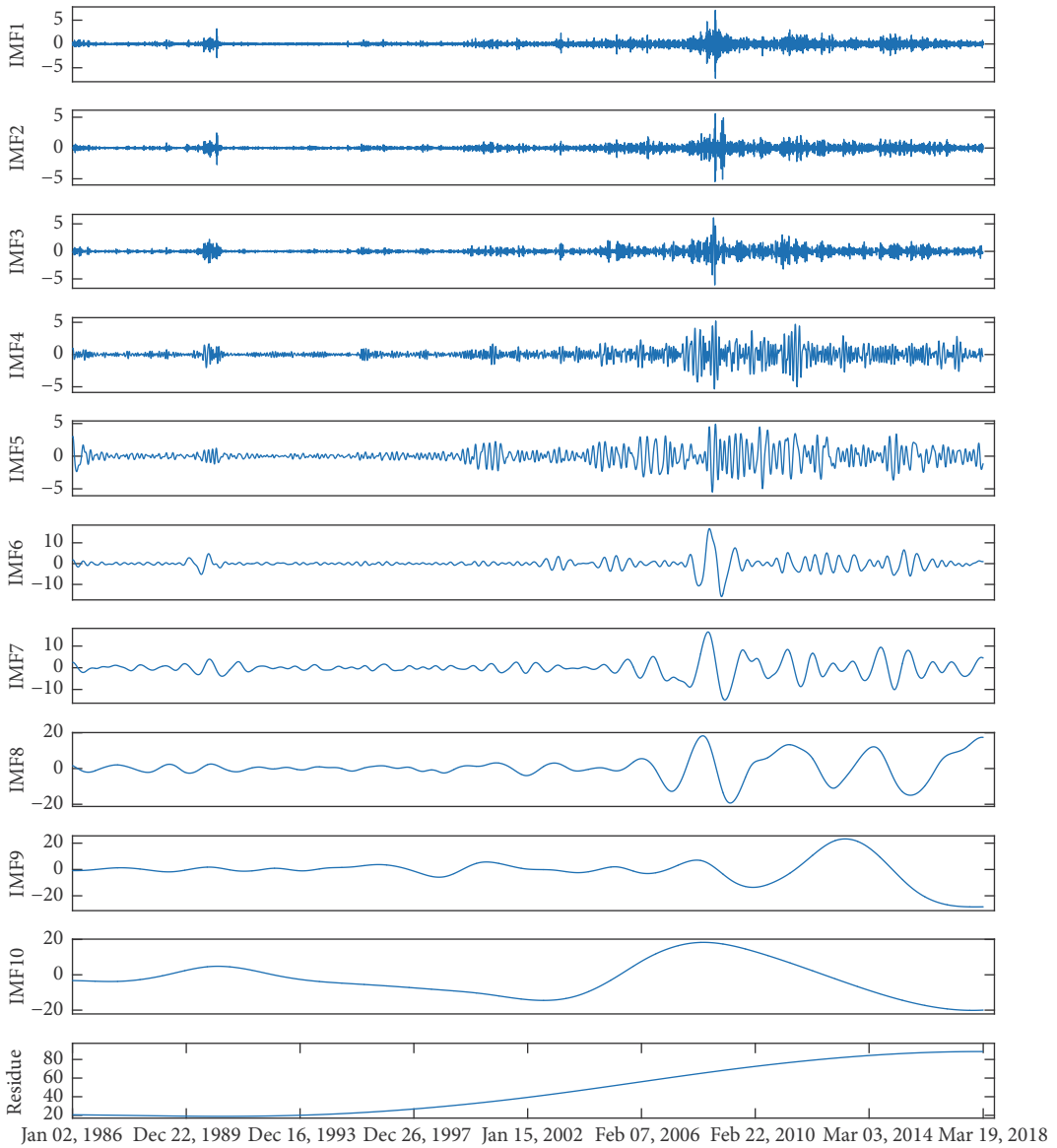


Figure 5: The IMFs and residue of WTI crude oil prices by CEEMDAN.

Table 4: MCS between XGBOOST, SVR, FNN, and ARIMA.

	HORIZON=1		HORIZON=3		HORIZON=6	
	T_R	T_{SQ}	T_R	T_{SQ}	T_R	T_{SQ}
XGBOOST	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
SVR	0.0004	0.0004	0.4132	0.4132	0.0200	0.0200
FNN	0.0002	0.0002	0.0248	0.0538	0.0016	0.0022
ARIMA	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

survival and the second best model on this horizon. When it comes to ARIMA, ARIMA almost performs as good as XGBOOST in terms of evaluation criteria of global errors but it does not pass the MCS. It indicates that ARIMA does not perform better than other models in most local errors of different samples.

4.4.2. Experimental Results of Ensemble Models. With EEMD or CEEMDAN, the results of forecasting the crude oil prices by XGBOOST, SVR, and FNN with horizons 1, 3, and 6, are shown in Table 5.

It can be seen from Table 5 that the RMSE and MAE values of the CEEMDAN-XGBOOST are the lowest ones

Table 5: The RMSE, MAE, and Dstat by the models with EEMD or CEEMDAN.

Horizon	Model	RMSE	MAE	Dstat
1	CEEMDAN-XGBOOST	**0.4151**	**0.3023**	0.8783
	EEMD-XGBOOST	0.9941	0.7685	0.7109
	CEEMDAN-SVR	0.8477	0.7594	**0.9054**
	EEMD-SVR	1.1796	0.9879	0.8727
	CEEMDAN-FNN	1.2574	1.0118	0.7597
	EEMD-FNN	2.6835	1.9932	0.7361
3	CEEMDAN-XGBOOST	**0.8373**	**0.6187**	0.6914
	EEMD-XGBOOST	1.4007	1.0876	0.6320
	CEEMDAN-SVR	1.2399	1.0156	**0.7092**
	EEMD-SVR	1.2366	1.0275	**0.7092**
	CEEMDAN-FNN	1.2520	0.9662	0.7061
	EEMD-FNN	1.2046	0.8637	0.6959
6	CEEMDAN-XGBOOST	**1.2882**	**0.9831**	0.6196
	EEMD-XGBOOST	1.7719	1.3765	0.6165
	CEEMDAN-SVR	1.3453	1.0296	**0.6683**
	EEMD-SVR	1.3730	1.1170	0.6485
	CEEMDAN-FNN	1.8024	1.3647	0.6422
	EEMD-FNN	2.7786	2.0495	0.6337

among those by all methods with each horizon. For example, with horizon 1, the values of RMSE and MAE are 0.4151 and 0.3023, which are far less than the second values of RMSE and MAE, i.e., 0.8477 and 0.7594, respectively. With the horizon increases, the corresponding values of each model increase, in terms of RMSE and MAE. However, the CEEMDAN-XGBOOST still achieves the lowest values of RMSE and MAE with each horizon. Regarding the values of Dstat, all the values are far greater than those by random guesses, showing that the "decomposition and ensemble" framework is effective for directional forecasting. Specifically, the values of Dstat are in the range between 0.6165 and 0.9054. The best Dstat values in all horizons are achieved by CEEMDAN-SVR or EEMD-SVR, showing that, among the forecasters, SVR is the best one for directional forecasting, although corresponding values of RMSE and MAE are not the best. As for the decomposition methods, when the forecasters are fixed, CEEMDAN outperforms EEMD in 8, 8, and 8 out of 9 cases in terms of RMSE, MAE, and Dstat, respectively, showing the advantages of CEEMDAN over EEMD. Regarding the forecasters, when combined with CEEMDAN, XGBOOST is always superior to other forecasters in terms of RMSE and MAE. However, when combined with EEMD, XGBOOST outperforms SVR and FNN with horizon 1, and FNN with horizon 6 in terms of RMSE and MAE. With horizons 1 and 6, FNN achieves the worst results of RMSE and MAE. The results also show that good values of RMSE usually are associated with good values of MAE. However, good values of RMSE or MAE do not always mean good Dstat directly.

For the ensemble models, we also took a Wilcoxon signed rank test and an MCS test based on the errors of pairs of models. We set 0.2 as the level of significance in MCS, and 0.05 as the level of significance in WSRT. The results are shown in Tables 6 and 7.

From these two tables, it can be seen that, regarding the results of WSRT, the p-value between CEEMDAN-XGBOOST and any other models except EEMD-FNN are below 0.05, demonstrating that there is a significant difference on the population mean ranks between CEEMDAN-XGBOOST and any other models except EEMD-FNN. What is more, the MCS shows that the p-value of T_R and T_{SQ} of CEEMDAN-XGBOOST is always equal to 1.000 and demonstrates that CEEMDAN-XGBOOST is the optimal model among all models in terms of global errors and local errors. Meanwhile, the p-values of T_R and T_{SQ} of EEMD-FNN are greater than other models except CEEMDAN-XGBOOST and become the second best model with horizons 3 and 6 in MCS. Meanwhile, with the horizon 6, the CEEMDAN-SVR is also the second best model. Besides, the p-values of T_R and T_{SQ} of EEMD-SVR and CEEMDAN-SVR are up to 0.2 and they become the surviving models with horizon 6 in MCS.

From the results of single models and ensemble models, we can draw the following conclusions: (1) single models usually cannot achieve satisfactory results, due to the nonlinearity and nonstationarity of raw crude oil prices. As a single forecaster, XGBOOST can achieve slightly better results than some state-of-the-art algorithms; (2) ensemble models can significantly improve the forecasting accuracy in terms of several evaluation criteria, following the "decomposition and ensemble" framework; (3) as a decomposition method, CEEMDAN outperforms EEMD in most cases; (4) the extensive experiments demonstrate that the proposed CEEMDAN-XGBOOST is promising for forecasting crude oil prices.

4.5. Discussion. In this subsection, we will study the impacts of several parameters related to the proposed CEEMDAN-XGBOOST.

4.5.1. The Impact of the Number of Realizations in CEEMDAN. In (2), it is shown that there are N realizations $x^i[t]$ in CEEMDAN. We explore how the number of realizations in CEEMDAN can influence on the results of forecasting crude oil prices by CEEMDAN-XGBOOST with horizon 1 and lag 6. And we set the number of realizations in CEEMDAN in the range of {10,25,50,75,100,250,500,750,1000}. The results are shown in Figure 6.

It can be seen from Figure 6 that, for RMSE and MAE, the bigger the number of realizations is, the more accurate results the CEEMDAN-XGBOOST can achieve. When the number of realizations is less than or equal to 500, the values of both RMSE and MAE decrease with increasing of the number of realizations. However, when the number is greater than 500, these two values are increasing slightly. Regarding Dstat, when the number increases from 10 to 25, the value of Dstat increases rapidly, and then it increases slowly with the number increasing from 25 to 500. After that, Dstat decreases slightly. It is shown that the value of Dstat reaches the top values with the number of realizations 500. Therefore, 500 is the best value for the number of realizations in terms of RMSE, MAE, and Dstat.

Table 6: Wilcoxon signed rank test between XGBOOST, SVR, and FNN with EEMD or CEEMDAN.

	CEEMDAN-XGBOOST	EEMD-XGBOOST	CEEMDAN-SVR	EEMD-SVR	CEEMDAN-FNN	EEMD-FNN
CEEMDAN-XGBOOST	1	3.5544e-05	1.8847e-50	0. 0028	1.6039e-187	0.0726
EEMD-XGBOOST	3.5544e-05	1	4.5857e-07	0. 3604	8.2912e-82	0.0556
CEEMDAN-SVR	1.8847e-50	4.5857e-07	1	4.9296e-09	5.7753e-155	8.6135e-09
EEMD-SVR	0.0028	0. 3604	4.9296e-09	1	2.5385e-129	0.0007
CEEMDAN-FNN	1.6039e-187	8.2912e-82	5.7753e-155	2.5385e-129	1	8.1427e-196
EEMD-FNN	0.0726	0.0556	8.6135e-09	0.0007	8.1427e-196	1

Table 7: MCS between XGBOOST, SVR, and FNN with EEMD or CEEMDAN.

	HORIZON=1		HORIZON=3		HORIZON=6	
	T_R	T_{SQ}	T_R	T_{SQ}	T_R	T_{SQ}
CEEMDAN-XGBOOST	1	1	1	1	1	1
EEMD-XGBOOST	0	0	0	0	0	0.0030
CEEMDAN-SVR	0	0.0002	0.0124	0.0162	0. 8268	0.8092
EEMD-SVR	0	0	0.0008	0.004	0.7872	0.7926
CEEMDAN-FNN	0	0	0.0338	0.0532	0.2924	0.3866
EEMD-FNN	0	0.0002	0.4040	0.4040	0. 8268	0.8092

4.5.2. The Impact of the Lag Orders. In this section, we explore how the number of lag orders impacts the prediction accuracy of CEEMDAN-XGBOOST on the horizon of 1. In this experiment, we set the number of lag orders from 1 to 10, and the results are shown in Figure 7.

According to the empirical results shown in Figure 7, it can be seen that as the lag order increases from 1 to 2, the values of RMSE and MAE decrease sharply while that of Dstat increases drastically. After that, the values of RMSE of MAE remain almost unchanged (or increase very slightly) with the increasing of lag orders. However, for Dstat, the value increases sharply from 1 to 2 and then decreases from 2 to 3. After the lag order increases from 3 to 5, the Dstat stays almost stationary. Overall, when the value of lag order is up to 5, it reaches a good tradeoff among the values of RMSE, MAE, and Dstat.

4.5.3. The Impact of the Noise Strength in CEEMDAN. Apart from the number of realizations, the noise strength in CEEMDAN, which stands for the standard deviation of the white noise in CEEMDAN, also affects the performance of CEEMDAN-XGBOOST. Thus, we set the noise strength in the set of {0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07} to explore how the noise strength in CEEMDAN affects the prediction accuracy of CEEMDAN-XGBOOST on a fixed horizon 1 and a fixed lag 6.

As shown in Figure 8, when the noise strength in CEEMDAN is equal to 0.05, the values of RMSE, MAE and Dstat achieve the best results simultaneously. When the noise strength is less than or equal to 0.05 except 0.03, the values of RMSE, MAE and Dstat become better and better with the increase of the noise strength. However, when the strength is greater than 0.05, the values of RMSE, MAE and Dstat become worse and worse. The figure indicates that the noise strength has a great impact on forecasting results, and an ideal range for it is about 0.04-0.06.

5. Conclusions

In this paper, we propose a novel model, namely, CEEMDAN-XGBOOST, to forecast crude oil prices. At first, CEEMDAN-XGBOOST decomposes the sequence of crude oil prices into several IMFs and a residue with CEEMDAN. Then, it forecasts the IMFs and the residue with XGBOOST individually. Finally, CEEMDAN-XGBOOST computes the sum of the prediction of the IMFs and the residue as the final forecasting results. The experimental results show that the proposed CEEMDAN-XGBOOST significantly outperforms the compared methods in terms of RMSE and MAE. Although the performance of the CEEMDAN-XGBOOST on forecasting the direction of crude oil prices is not the best, the MCS shows that the CEEMDAN-XGBOOST is still the optimal model. Meanwhile, it is proved that the number of the realizations, lag, and the noise strength for CEEMDAN are the vital factors which have great impacts on the performance of the CEEMDAN-XGBOOST.

In the future, we will study the performance of the CEEMDAN-XGBOOST on forecasting crude oil prices with different periods. We will also apply the proposed approach for forecasting other energy time series, such as wind speed, electricity load, and carbon emissions prices.

Data Availability

The data used to support the findings of this study are included within the article.

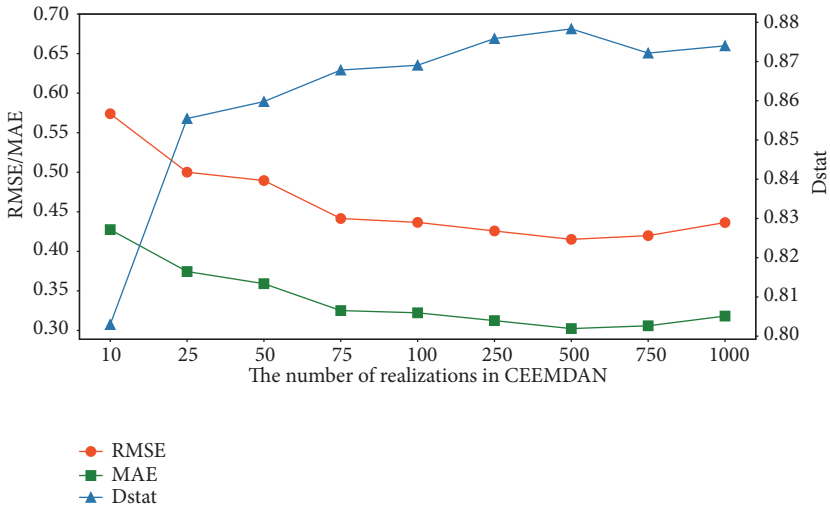


Figure 6: The impact of the number of IMFs with CEEMDAN-XGBOOST.

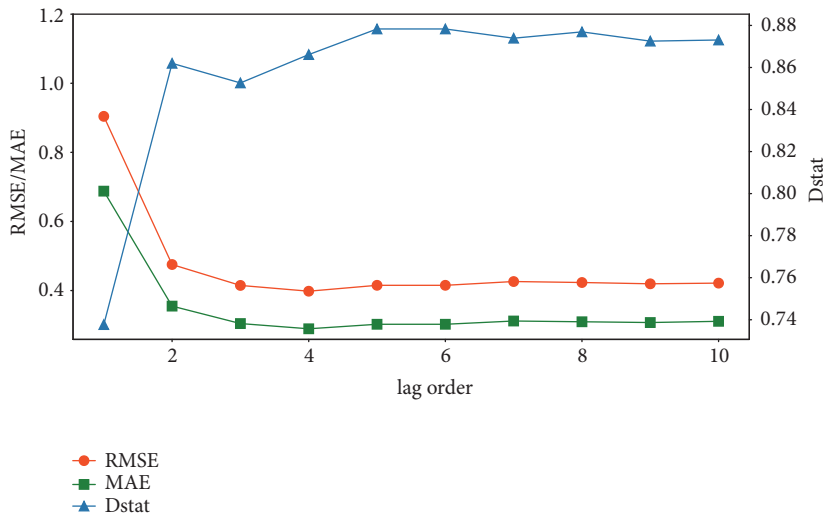


Figure 7: The impact of the number of lag orders with CEEMDAN-XGBOOST.

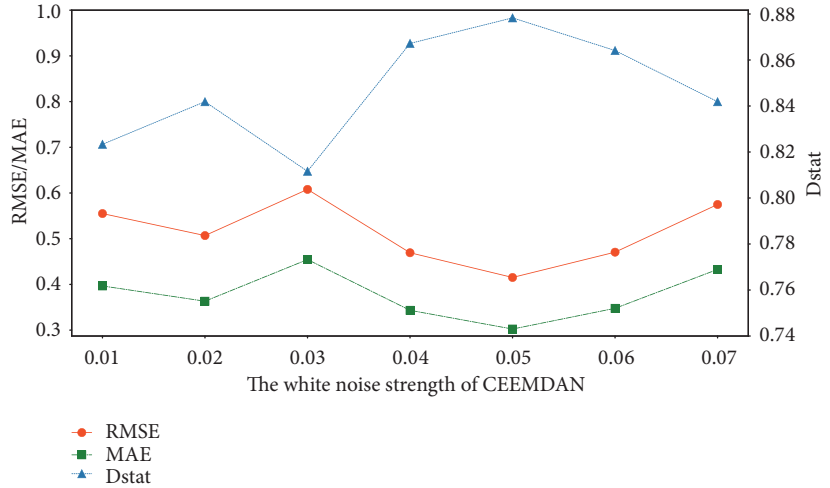


Figure 8: The impact of the value of the white noise of CEEMDAN with CEEMDAN-XGBOOST.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the Fundamental Research Funds for the Central Universities (Grants no. JBK1902029, no. JBK1802073, and no. JBK170505), the Natural Science Foundation of China (Grant no. 71473201), and the Scientific Research Fund of Sichuan Provincial Education Department (Grant no. 17ZB0433).

References

[1] H. Miao, S. Ramchander, T. Wang, and D. Yang, "Influential factors in crude oil price forecasting," *Energy Economics*, vol. 68, pp. 77–88, 2017.

[2] L. Yu, S. Wang, and K. K. Lai, "Forecasting crude oil price with an EMD-based neural network ensemble learning paradigm," *Energy Economics*, vol. 30, no. 5, pp. 2623–2635, 2008.

[3] M. Ye, J. Zyren, C. J. Blumberg, and J. Shore, "A short-run crude oil price forecast model with ratchet effect," *Atlantic Economic Journal*, vol. 37, no. 1, pp. 37–50, 2009.

[4] C. Morana, "A semiparametric approach to short-term oil price forecasting," *Energy Economics*, vol. 23, no. 3, pp. 325–338, 2001.

[5] H. Naser, "Estimating and forecasting the real prices of crude oil: A data rich model using a dynamic model averaging (DMA) approach," *Energy Economics*, vol. 56, pp. 75–87, 2016.

[6] X. Gong and B. Lin, "Forecasting the good and bad uncertainties of crude oil prices using a HAR framework," *Energy Economics*, vol. 67, pp. 315–327, 2017.

[7] F. Wen, X. Gong, and S. Cai, "Forecasting the volatility of crude oil futures using HAR-type models with structural breaks," *Energy Economics*, vol. 59, pp. 400–413, 2016.

[8] H. Chiroma, S. Abdul-Kareem, A. Shukri Mohd Noor et al., "A Review on Artificial Intelligence Methodologies for the Forecasting of Crude Oil Price," *Intelligent Automation and Soft Computing*, vol. 22, no. 3, pp. 449–462, 2016.

[9] S. Wang, L. Yu, and K. K. Lai, "A novel hybrid AI system framework for crude oil price forecasting," in *Data Mining and Knowledge Management*, Y. Shi, W. Xu, and Z. Chen, Eds., vol. 3327 of *Lecture Notes in Computer Science*, pp. 233–242, Springer, Berlin, Germany, 2004.

[10] J. Baruník and B. Malinská, "Forecasting the term structure of crude oil futures prices with neural networks," *Applied Energy*, vol. 164, pp. 366–379, 2016.

[11] Y. Chen, K. He, and G. K. F. Tso, "Forecasting crude oil prices-a deep learning based model," *Procedia Computer Science*, vol. 122, pp. 300–307, 2017.

[12] R. Tehrani and F. Khodayar, "A hybrid optimized artificial intelligent model to forecast crude oil using genetic algorithm," *African Journal of Business Management*, vol. 5, no. 34, pp. 13130–13135, 2011.

[13] L. Yu, Y. Zhao, and L. Tang, "A compressed sensing based AI learning paradigm for crude oil price forecasting," *Energy Economics*, vol. 46, no. C, pp. 236–245, 2014.

[14] L. Yu, H. Xu, and L. Tang, "LSSVR ensemble learning with uncertain parameters for crude oil price forecasting," *Applied Soft Computing*, vol. 56, pp. 692–701, 2017.

[15] Y.-L. Qi and W.-J. Zhang, "The improved SVM method for forecasting the fluctuation of international crude oil price," in *Proceedings of the 2009 International Conference on Electronic Commerce and Business Intelligence (ECBI '09)*, pp. 269–271, IEEE, China, June 2009.

[16] M. S. AL-Musaylh, R. C. Deo, Y. Li, and J. F. Adamowski, "Two-phase particle swarm optimized-support vector regression hybrid model integrated with improved empirical mode decomposition with adaptive noise for multiple-horizon electricity demand forecasting," *Applied Energy*, vol. 217, pp. 422–439, 2018.

[17] L. Huang and J. Wang, "Forecasting energy fluctuation model by wavelet decomposition and stochastic recurrent wavelet neural network," *Neurocomputing*, vol. 309, pp. 70–82, 2018.

[18] H. Zhao, M. Sun, W. Deng, and X. Yang, "A new feature extraction method based on EEMD and multi-scale fuzzy entropy for motor bearing," *Entropy*, vol. 19, no. 1, Article ID 14, 2017.

[19] W. Deng, S. Zhang, H. Zhao, and X. Yang, "A Novel Fault Diagnosis Method Based on Integrating Empirical Wavelet Transform and Fuzzy Entropy for Motor Bearing," *IEEE Access*, vol. 6, pp. 35042–35056, 2018.

[20] Q. Fu, B. Jing, P. He, S. Si, and Y. Wang, "Fault Feature Selection and Diagnosis of Rolling Bearings Based on EEMD and Optimized Elman_AdaBoost Algorithm," *IEEE Sensors Journal*, vol. 18, no. 12, pp. 5024–5034, 2018.

[21] T. Li and M. Zhou, "ECG classification using wavelet packet entropy and random forests," *Entropy*, vol. 18, no. 8, p. 285, 2016.

[22] M. Blanco-Velasco, B. Weng, and K. E. Barner, "ECG signal denoising and baseline wander correction based on the empirical mode decomposition," *Computers in Biology and Medicine*, vol. 38, no. 1, pp. 1–13, 2008.

[23] J. Lee, D. D. McManus, S. Merchant, and K. H. Chon, "Automatic motion and noise artifact detection in holter ECG data using empirical mode decomposition and statistical approaches," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 6, pp. 1499–1506, 2012.

[24] L. Fan, S. Pan, Z. Li, and H. Li, "An ICA-based support vector regression scheme for forecasting crude oil prices," *Technological Forecasting & Social Change*, vol. 112, pp. 245–253, 2016.

[25] E. Jianwei, Y. Bao, and J. Ye, "Crude oil price analysis and forecasting based on variational mode decomposition and independent component analysis," *Physica A: Statistical Mechanics and Its Applications*, vol. 484, pp. 412–427, 2017.

[26] N. E. Huang, Z. Shen, S. R. Long et al., "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 454, pp. 903–995, 1998.

[27] Z. Wu and N. E. Huang, "Ensemble empirical mode decomposition: a noise-assisted data analysis method," *Advances in Adaptive Data Analysis*, vol. 1, no. 1, pp. 1–41, 2009.

[28] L. Yu, W. Dai, L. Tang, and J. Wu, "A hybrid grid-GA-based LSSVR learning paradigm for crude oil price forecasting," *Neural Computing and Applications*, vol. 27, no. 8, pp. 2193–2215, 2016.

[29] L. Tang, W. Dai, L. Yu, and S. Wang, "A novel CEEMD-based eelm ensemble learning paradigm for crude oil price forecasting," *International Journal of Information Technology & Decision Making*, vol. 14, no. 1, pp. 141–169, 2015.

[30] T. Li, M. Zhou, C. Guo et al., "Forecasting crude oil price using EEMD and RVM with adaptive PSO-based kernels," *Energies*, vol. 9, no. 12, p. 1014, 2016.

[31] T. Li, Z. Hu, Y. Jia, J. Wu, and Y. Zhou, "Forecasting Crude Oil Prices Using Ensemble Empirical Mode Decomposition and Sparse Bayesian Learning," *Energies*, vol. 11, no. 7, 2018.

[32] M. E. Torres, M. A. Colominas, G. Schlotthauer, and P. Flandrin, "A complete ensemble empirical mode decomposition with adaptive noise," in *Proceedings of the 36th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '11)*, pp. 4144–4147, IEEE, Prague, Czech Republic, May 2011.

[33] M. A. Colominas, G. Schlotthauer, and M. E. Torres, "Improved complete ensemble EMD: a suitable tool for biomedical signal processing," *Biomedical Signal Processing and Control*, vol. 14, no. 1, pp. 19–29, 2014.

[34] T. Peng, J. Zhou, C. Zhang, and Y. Zheng, "Multi-step ahead wind speed forecasting using a hybrid model based on two-stage decomposition technique and AdaBoost-extreme learning machine," *Energy Conversion and Management*, vol. 153, pp. 589–602, 2017.

[35] S. Dai, D. Niu, and Y. Li, "Daily peak load forecasting based on complete ensemble empirical mode decomposition with adaptive noise and support vector machine optimized by modified grey Wolf optimization algorithm," *Energies*, vol. 11, no. 1, 2018.

[36] Y. Lv, R. Yuan, T. Wang, H. Li, and G. Song, "Health Degradation Monitoring and Early Fault Diagnosis of a Rolling Bearing Based on CEEMDAN and Improved MMSE," *Materials* , vol. 11, no. 6, p. 1009, 2018.

[37] R. Abdelkader, A. Kaddour, A. Bendiabdellah, and Z. Derouiche, "Rolling Bearing Fault Diagnosis Based on an Improved Denoising Method Using the Complete Ensemble Empirical Mode Decomposition and the Optimized Thresholding Operation," *IEEE Sensors Journal*, vol. 18, no. 17, pp. 7166–7172, 2018.

[38] Y. Lei, Z. Liu, J. Ouazri, and J. Lin, "A fault diagnosis method of rolling element bearings based on CEEMDAN," *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, vol. 231, no. 10, pp. 1804–1815, 2017.

[39] T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2016*, pp. 785–794, ACM, New York, NY, USA, August 2016.

[40] B. Zhai and J. Chen, "Development of a stacked ensemble model for forecasting and analyzing daily average PM2.5 concentrations in Beijing, China," *Science of the Total Environment*, vol. 635, pp. 644–658, 2018.

[41] S. P. Chatzis, V. Siakoulis, A. Petropoulos, E. Stavroulakis, and N. Vlachogiannakis, "Forecasting stock market crisis events using deep and statistical machine learning techniques," *Expert Systems with Applications*, vol. 112, pp. 353–371, 2018.

[42] J. Ke, H. Zheng, H. Yang, and X. Chen, "Short-term forecasting of passenger demand under on-demand ride services: A spatio-temporal deep learning approach," *Transportation Research Part C: Emerging Technologies*, vol. 85, pp. 591–608, 2017.

[43] J. Friedman, T. Hastie, and R. Tibshirani, "Special Invited Paper. Additive Logistic Regression: A Statistical View of Boosting: Rejoinde," *The Annals of Statistics*, vol. 28, no. 2, pp. 400–407, 2000.

[44] J. D. Gibbons and S. Chakraborti, "Nonparametric statistical inference," in *International Encyclopedia of Statistical Science*, M. Lovric, Ed., pp. 977–979, Springer, Berlin, Germany, 2011.

[45] P. R. Hansen, A. Lunde, and J. M. Nason, "The model confidence set," *Econometrica*, vol. 79, no. 2, pp. 453–497, 2011.

[46] H. Liu, H. Q. Tian, and Y. F. Li, "Comparison of two new ARIMA-ANN and ARIMA-Kalman hybrid methods for wind speed prediction," *Applied Energy*, vol. 98, no. 1, pp. 415–424, 2012.

Hindawi
Complexity
Volume 2019, Article ID 2782715, 14 pages
https://doi.org/10.1155/2019/2782715

Research Article

Development of Multidecomposition Hybrid Model for Hydrological Time Series Analysis

Hafiza Mamona Nazir[1], **Ijaz Hussain**[1], **Muhammad Faisal**[2,3], **Alaa Mohamd Shoukry**[4,5], **Showkat Gani**[6] and **Ishfaq Ahmad**[7,8]

[1]*Department of Statistics, Quaid-i-Azam University, Islamabad, Pakistan*
[2]*Faculty of Health Studies, University of Bradford, Bradford BD7 1DP, UK*
[3]*Bradford Institute for Health Research, Bradford Teaching Hospitals NHS Foundation Trust, Bradford, UK*
[4]*Arriyadh Community College, King Saud University, Riyadh, Saudi Arabia*
[5]*KSA Workers University, El-Mansoura, Egypt*
[6]*College of Business Administration, King Saud University, Al-Muzahimiyah, Saudi Arabia*
[7]*Department of Mathematics, College of Science, King Khalid University, Abha 61413, Saudi Arabia*
[8]*Department of Mathematics and Statistics, Faculty of Basic and Applied Sciences, International Islamic University, 44000 Islamabad, Pakistan*

Correspondence should be addressed to Ijaz Hussain; ijaz@qau.edu.pk

Received 1 October 2018; Accepted 13 December 2018; Published 2 January 2019

Guest Editor: Pedro Palos

Accurate prediction of hydrological processes is key for optimal allocation of water resources. In this study, two novel hybrid models are developed to improve the prediction precision of hydrological time series data based on the principal of three stages as denoising, decomposition, and decomposed component prediction and summation. The proposed architecture is applied on daily rivers inflow time series data of Indus Basin System. The performances of the proposed models are compared with traditional single-stage model (without denoised and decomposed), the hybrid two-stage model (with denoised), and existing three-stage hybrid model (with denoised and decomposition). Three evaluation measures are used to assess the prediction accuracy of all models such as Mean Relative Error (MRE), Mean Absolute Error (MAE), and Mean Square Error (MSE). The proposed, three-stage hybrid models have shown improvement in prediction accuracy with minimum MRE, MAE, and MSE for all case studies as compared to other existing one-stage and two-stage models. In summary, the accuracy of prediction is improved by reducing the complexity of hydrological time series data by incorporating the denoising and decomposition.

1. Introduction

Accurate prediction of hydrological processes is key for optimal allocation of water resources. It is challenging because of its nonstationary and multiscale stochastic characteristics of hydrological process which are affected not only by climate change but also by other socioeconomic development projects. The human activities also effected the climate change through contributing in Earth's atmosphere by burning of fossil fuels which release carbon dioxide in atmosphere. Instead of these, greenhouse and aerosols have made effect on Earth's atmosphere by altering in-out coming solar radiations which is the part of Earth's energy balance. This makes the prediction of hydrological time series data challenging. To predict such hydrological processes, two broad types of models are commonly used, one is the process-based models which further included the lumped conceptual models, hydrological model, and one-two-dimensional hydrodynamic models [1], and the second is data driven models which included autoregressive moving averages and artificial neural network (which are also known as black box models). The process-based models considered the physical mechanism of stochastic hydrological processes, which requires a large amount of data for calibration and validation

[2]. Moreover, physical-based models demand the scientific principles of energy and water movements spatiotemporally. Zahidul [3] concluded that unavailability of sufficient amount of data and scientific knowledge of water movement can lead to poor understanding of hydrological system which makes the hydrological modeling a challenging task. In order to overcome these drawbacks, hydrologists used data driven models to efficiently model the hydrological process [4, 5]. The data driven models only take the advantage of inherent the input-output relationship through data manipulation without considering the internal physical process. The data-driven models are efficient over the process-driven models by appreciating the advantage of less demanding the quantitative data, simple formulation with better prediction performance [6]. These data-driven models are further divided into two categories: simple traditional statistical techniques and more complex machine learning methods. In the last few decades, many traditional statistical time series models are developed including Autoregressive (AR), Moving Averages (MA), Autoregressive Moving Averages (ARMA), and Autoregressive Integrated Moving Averages (ARIMA) [7]. Application of ARIMA model to monitoring hydrological processes like river discharge is common and successfully applied [8]. But the problem with all these traditional statistical techniques required that the time series data to be stationary. However, hydrological data was characterized as both nonstationary and nonlinear due to its time varying nature. Therefore, these techniques are not enough to capture the nonlinear characteristics of hydrological series [6]. To rescue the drawbacks of existing traditional models, machine learning (ML) algorithms have been put forward and widely exploited, which provide powerful solution to the instability of hydrological time series data [4]. ML models include Artificial Neural Network (ANN), Support Vector Machine (SVM), and random forest and genetic algorithms [9–14]. Riad et al. [5] developed an ANN to model the nonlinear relation between rainfall and runoff and concluded that ANN model is better to model the complex hydrological system over the traditional statistical models. However, these ML methods have their own drawbacks such as overfitting and being sensitive to parameter selection. In addition, there are two main drawbacks of using ML models: first is that ML models ignore the time varying characteristics of hydrological time series data and secondly the hydrological data contains noises which deprive the researchers to accurately predict the hydrological time series data in an effective way [15]. These time varying and noise corrupted characteristics of hydrological time series data require hybrid approaches to model the complex characteristics of hydrological time series data [16].

To conquer the limitations of existing single models, some hybrid algorithms such as data preprocessing methods are utilized with data-driven models with the hope to enhance the prediction performance of complex hydrological time series data by extracting time varying components with noise reduction. These preprocess based hybrid models have already been applied in hydrology [2]. The framework of hybrid model usually comprised "decomposition," "prediction," and "ensemble" [2, 6, 13]. The most commonly used data preprocessing method is wavelet analysis (WA) which is used to decompose the nonlinear and nonstationary hydrological data into multiscale components [13]. These processed multiscale components are further used as inputs in black box models at prediction stage and finally predicted components are ensemble to get final predictions. Peng et al. [6] proposed hybrid model by using empirical wavelet transform and ANN for reliable stream flow forecasting. They demonstrated their proposed hybrid model efficiency over single models. Later, Wu et al. [11] exploited a two-stage hybrid model by incorporating Wavelet Multi-Resolution Analysis (WMRA), and other data preprocessing methods as MA, singular spectrum analysis with ANN to enhance the estimate of daily flows. They proposed five models including ANN-MA, ANN-SSA1, ANN-SSA2, ANN-WMRA1, and ANN-WMRA2 and suggested that decomposition with MA model performs better than WMRA. An improvement in wavelet decomposition method has been made to get more accurate hybrid results comprising WA [17]. However, the problem which reduces the performance of WA, i.e., selection of mother wavelet basis function, is still an open debate as the selection of mother wavelet is subjectively determined among many wavelet basis functions. The optimality of multiscale characteristics entirely depends on the choice of mother wavelet function as poorly selected mother wavelet function can lead to more uncertainty in time-scale components. To overcome this drawback, Huang et al. [18] proposed a purely data-driven Empirical Mode Decomposition (EMD) technique. The objective of EMD is to decompose the nonlinear and nonstationary data adaptively into number of oscillatory components called Intrinsic Mode Decomposition (IMF). A number of studies have been conducted combining the EMD with data driven models [15, 18–21]. Specifically in hydrology, EMD is used with ANN for wind speed and stream flow prediction [15, 20]. Agana and Homaifar [21] developed the EMD-based predictive deep belief network for accurately predicting and forecasting the Standardized Stream flow Index (SSI). Their study manifested that their proposed model is better than the existing standard methods with the improvement in prediction accuracy of SSI. However, Kang et al. [22] revealed that EMD suffers with mode mixing problem which ultimately affects the efficiency of decomposition. In order to deal with this mode mixing problem, Wu and Hang [23] proposed an improved EMD by successively introducing white Gauss noise in signals, called Ensemble Empirical Mode Decomposition (EEMD) that addresses the issue of frequent apparent of mode mixing in EMD. Later, EEMD was effectively used as data decomposition method to extract the multiscale characteristics [24–26]. Di et al. [2] proposed a four-stage hybrid model (based on EEMD for decomposition) to improve the prediction accuracy by reducing the redundant noises and concluded that coupling the appropriate data decomposition with EEMD method with data driven models could improve the prediction performance compared to existing EMD based hybrid model. Jiang et al. [26] proposed another two-stage hybrid approach coupling EEMD with data-driven models to forecast high speed rail passenger flow to estimate daily ridership. They suggested that their proposed hybrid model is more suitable

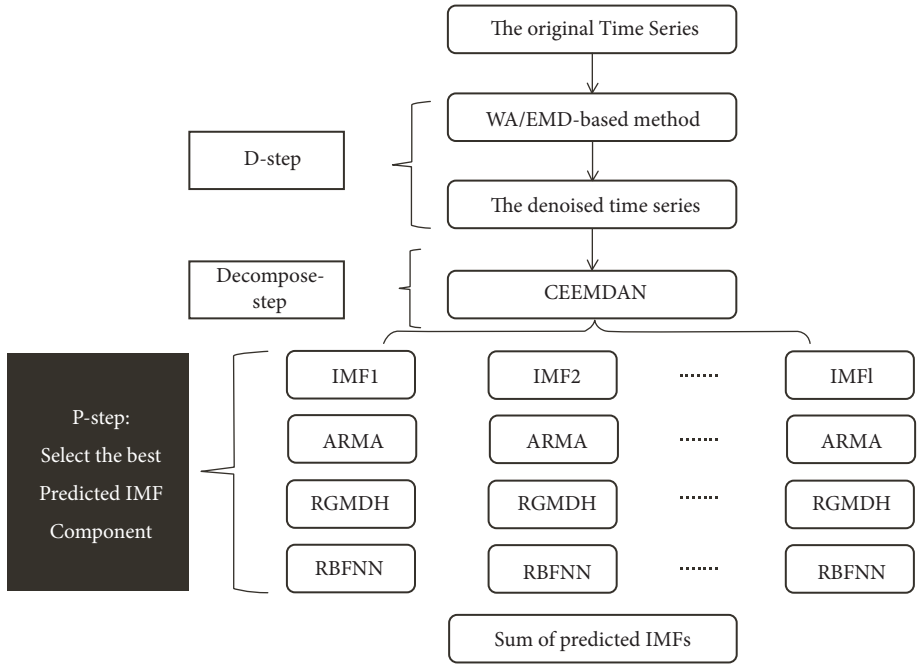


Figure 1: The proposed WA/EMD-CEEMDAN-MM structure to predict hydrological time series data.

for short term prediction by accounting for the day to day variation over other hybrid and single models. However, due to successive addition of independent white Gauss noise, the performance of EEMD is affected which reduces the accuracy of extracted IMFs through EEMD algorithm. Dai et al. [27] reported in their study that EEMD based hybrid models did not perform appropriately due to independent noise addition.

This study aimed to develop a robust hybrid model to decompose the hydrological time varying characteristics using CEEMDAN [28]. The CEEMDAN successively adds white noise, following the steps of EEMD, with interference in each decomposition level to overcome the drawback of EEMD algorithm. Dai et al. [27] developed a model comprising CEEMDAN to daily peak load forecasting which shows robust decomposed ability for reliable forecasting. Therefore, the purpose of using CEEMDAN method for decomposition in this study is to find an effective way to decompose the nonlinear data which enhances the prediction accuracy of the complex hydrological time series data [27].

2. Proposed Methods

In this study, two novel approaches are proposed to enhance the prediction accuracy of the hydrological time series. Both models have the same layout except in stage of denoising, where two different approaches have been used to remove noises from hydrological time series data. In both models, at decomposition stage, an improved version of EEMD, i.e., CEEMDAN, is used to find oscillations, i.e., the high to low frequencies in terms of IMF. At prediction stage, multi-models are used to accurately predict the extracted IMFs by considering the nature of IMFs instead of using only single stochastic model. The purpose of using multimodel is two-way: one is for accurately predicting the IMFs by considering the nature of IMFs and the other is to assess the performance of simple and complex models after reducing the complexity of hydrological time series data through decomposition. Predicted IMFs are added to get the final prediction of hydrological time series. The proposed three stages involve denoising (D-step), decomposition (Decompose-step), and component prediction (P-step), which are briefly described below:

(1) **D-step**: WA and EMD based denoising methods are presented to remove the noises from hydrological time series data.

(2) **Decomposed-step:** Using CEEMDAN, two separately denoised series are decomposed into k IMFs and one residual.

(3) **P-step:** The denoised-decomposed series into k IMFs and one residual are predicted with linear stochastic and nonlinear machine learning models. The model with the lowest error rate of prediction is selected by three performance evaluation measures. Finally the predicted results are added to get the final prediction.

For convenient, two proposed methods as named as EMD (denoising), CEEMDAN (decomposing), MM (multi-models) i.e. EMD-CEEMDAN-MM and WA (denoising), CEEMDAN (denoising) and MM (multi-models) i.e. WA-CEEMDAN-MM. The proposed architecture of WA/EMD-CEEMDAN-MM is given in Figure 1.

2.1. D-Step. In hydrology time series data, noises or stochastic volatiles are inevitable component which ultimately reduced

the performance of prediction. To reduce the noise from data, many algorithms have been proposed in literature such as Fourier analysis, spectral analysis, WA, and EMD [29], as besides decomposition, these techniques have the ability to remove the noises from data. However, the spectral and Fourier analysis only considered the linear and stationary signals, whereas WA and EMD have the ability to address the nonlinear and nonstationary data with better performance. In this study, WA- and EMD-based threshold are adopted to reduce the stochastic volatiles from the hydrological data.

(i) Wavelet analysis based denoising: in order to remove noises, discrete wavelet threshold method is recognized as powerful mathematical functions with hard and soft threshold. With the help of symlet 8 mother wavelet [30], hydrological time series data is decomposed into approximation and details coefficients with the following equations, respectively [31];

$$a_{j,k} = \sum_{k=0}^{2^{N-j}-1} 2^{-j/2} \emptyset \left(2^{-j}t - k\right) \tag{1}$$

and

$$d_{j,k} = \sum_{j=1}^{J} \sum_{k=0}^{2^{N-j}-1} 2^{-j/2} \varphi \left(2^{-j}t - k\right) \tag{2}$$

After estimating the approximation and details coefficients, threshold is calculated for each coefficient to remove noises. The energy of data is distributed only on few wavelet coefficients with high magnitude whereas most of the wavelet coefficients are noisiest with low magnitude. To calculate the noise free coefficient, hard and soft threshold rules are opted, which are listed as follows, respectively [31]:

$$d'_{j,k} = \begin{cases} d_{j,k} & \left|d_{j,k}\right| \geq T_j \\ 0 & \left|d_{j,k}\right| < T_j \end{cases} \tag{3}$$

and

$$d'_{j,k} = \begin{cases} \operatorname{sgn}\left(d_{j,k}\right)\left(\left|d_{j,k}\right| - T_j\right) & \left|d_{j,k}\right| \geq T_j \\ 0 & \left|d_{j,k}\right| < T_j \end{cases} \tag{4}$$

where T_i is the threshold calculated as $T_j = a\sqrt{2E_j \ln(N)}$, $j = 1, 2, \ldots, J$, where a is constant which takes the values between 0.4 and 1.4 with step of 0.1 and $E_j = median(|d_{j,k}|,\ k = 1, 2, \ldots, N)/0.6745$ is median deviation of all details. Then, the decomposed data is reconstructed using the noise free details and approximations using the following equation:

$$\begin{aligned} \widehat{x(t)} = &\sum_{k=0}^{2^{N-j}-1} a'_{j,k} 2^{-j/2} \emptyset \left(2^{-j}t - k\right) \\ &+ \sum_{j=1}^{J} \sum_{k=0}^{2^{N-j}-1} d'_{j,k} 2^{-j/2} \varphi \left(2^{-j}t - k\right) \end{aligned} \tag{5}$$

where $a'_{j,k}$ is threshold approximation coefficient and $d'_{j,k}$ is threshold detailed coefficient.

(ii) Empirical mode decomposition based denoising: an EMD is data-driven algorithm which has been recently proposed to decompose nonlinear and nonstationary data into several oscillatory modes [18]. Due to adaptive nature, EMD directly decomposes data into number of IMFs by satisfying two conditions as follows: **(a)** From complete data set, the number of zero crossings and the number of extremes must be equal or differ at most by one; **(b)** the mean value of the envelope which is smoothed, through cubic spline interpolation, based on the local maxima and minima should be zero at all points.

The EMD structure is defined as follows:

(1) Identify all local maxima and minima from time series $x(t)$, $(t = 1, 2, \ldots, N)$ and make upper envelope of maxima $e_{\max(t)}$ and lower envelope minima $e_{\min(t)}$ through cubic spline interpolation.

(2) Find the mean of upper and lower envelope $m(t) = (e_{\max(t)} + e_{\min(t)})/2$. Find the difference between original series and extracted mean as

$$h(t) = x(t) - m(t) \tag{6}$$

(3) Check the properties defined in **(a)** and **(b)** of $h(t)$; if both conditions are satisfied then mark this $h(t)$ as i^{th} IMF; then the next step will be to replace the original series by $r(t) = x(t) - h(t)$; if $h(t)$ is not IMF just replace $x(t)$ with $h(t)$.

(4) Repeat the process of (1-3), until the residue $r(t)$ becomes a monotone function from which no further IMFs can be extracted.

Finally, original series can be written as the sum of all extracted IMFs and residue as

$$x(t) = \sum_{i}^{m} h_i(t) + r(t) \tag{7}$$

where m is the number of IMFs, as $(i = 1, 2, \ldots, m)$ and $h_i(t)$ is the i^{th} IMF, and $r(t)$ is the trend of the signal. The way of denoised IMF is the same as mentioned in (3)-(5), except the last two IMFs which are used completely without denoising due to low frequencies. The subscript in EMD-based threshold case in (3)-(5) is replaced with i^{th} according to number of IMFs. The denoised signal is reconstructed as follows:

$$\widehat{x(t)} = \sum_{i=1}^{m-2} h_i(t) + \sum_{i=m-2}^{m} h_i(t) + r(t) \tag{8}$$

2.2. Decompose-Step: Decomposition Step. The EEMD method: the EEMD is a technique to stabilize the problem of mode mixing which arises in EMD and decomposes the nonlinear signals into number which contains the information of local time varying characteristics. The procedure of EEMD is as follows:

(a) Add a white Gaussian noise series to the original data set.

(b) Decompose the signals with added white noise into IMFs using conventional EMD method.

(c) Repeat steps (a) and (b) m^{th} time by adding different white noises ($m = 1, 2, \ldots, l$) in original signals.

(d) Obtain the ensemble means of all IMFs m^{th} ensemble time as the final results as $\overline{IMF}_k = \sum_{m=1}^{l} IMF_k^m / l$, where $k = 1, 2, \ldots, K$ is k^{th} IMF.

The CEEMDAN based decomposition: although the EEMD can reduce the mode mixing problem to some extent, due to the successive addition of white Gauss noise in EEMD, the error cannot be completely eliminated from IMFs. To overcome this situation, CEEMDAN function is introduced by Torres et al. [28]. We employed the CEEMDAN to decompose the hydrological time series data. The CEEMDAN is briefly described as follows:

(1) In CEEMDAN, extracted modes are defined as $\widetilde{IMF}_k$; in order to get complete decomposition we need to calculate the first residual by using the first $\widetilde{IMF}_1$, which is calculated by EEMD as $\widetilde{IMF}_1 = \sum_{m=1}^{l} IMF_1^m / l$.

(2) Then replace $x(t)$ by $r_1(t)$ where $r_1(t) = x(t) - \widetilde{IMF}_1$ and add white Gaussian noises, i.e., $w^m(t)$, m^{th} time in $r_1(t)$, and find the IMF by taking the average of first IMF to get the $\widetilde{IMF}_2$. Calculate $r_2(t) = r_1(t) - \widetilde{IMF}_2$ and repeat (2) until stoppage criteria are meet. However, selection of number of ensemble members and amplitude of white noise is still an open challenge but here in this paper the number of ensemble members is fixed as 100 and standard deviation of white noise is settled as 0.2.

(3) The resulting k^{th} decomposed modes, i.e., $\sum_{k=1}^{K} \widetilde{IMF}_k$, and one residual $R(t)$ are used for further prediction of hydrological time series.

More details of EMD, EEMD, and CEEMDAN are given in [20, 28].

2.3. P-Step

Prediction of All IMF's. In prediction stage, denoised IMFs are further used to predict the hydrological time series data as inputs by using simple stochastic and complex machine learning time series algorithms. The reason of using two types of model is that as first few IMFs contain high frequencies which are accurately predicted through complex ML models whenever, last IMFs contain low frequencies which are accurately predictable through simple stochastic models. The selected models are briefly described as follows.

The IMF prediction with ARIMA model: to predict the IMFs, autoregressive moving average model is used as follows:

$$IMF_t{}^i = \alpha_1 IMF_{t-1}{}^i + \cdots + \alpha_p IMF_{t-p}{}^i + \varepsilon_t{}^i + \beta_1 \varepsilon_{t-1}{}^i + \cdots + \beta_q \varepsilon_{t-q}{}^i \quad (9)$$

Table 1: Transfer functions of GMDH-NN algorithms.

Transfer Functions	
Sigmoid function	$z = \frac{1}{(1+e^{-y})}$
Tangent function	$z = \tan(y)$
Polynomial function	$z = y$
Radial basis function	$z = e^{-y^2}$

Here, $IMF_t{}^i$ is the i^{th} IMF and $\varepsilon_t{}^i$ is the i^{th} residual of CEEMDAN where p is autoregressive lag and q is moving average lag value. Often the case, time series is not stationary; [7] made a proposal that differencing to an appropriate degree can make the time series stationary; if this is the case then the model is said to be ARIMA (p, d, q) where d is the difference value which is used to make the series stationary.

The IMF prediction with group method of data handling type neural network: ANN has been proved to be a powerful tool to model complex nonlinear system. One of the submodels of NN, which is constructed to improve explicit polynomial model by self-organizing, is Group Method of Data Handling-type Neural Network (GMDH-NN) [32]. The GMDH-NN has a successful application in a diverse range of area; however, in hydrological modeling it is still scarce. The algorithm of GMDH-NN worked by considering the pairwise relationship between all selected lagged variables. Each selected combination of pairs entered in a neuron and output is constructed for each neuron. The structure of GMDH-NN is illustrated in Figure 2 with four variables, two hidden and one output layer. According to evaluation criteria, some neurons are selected as shown in Figure 2, four neurons are selected, and the output of these neurons becomes the input for next layer. A prediction mean square criterion is used for neuron output selection. The process is continued till the last layer. In the final layer, only single best predicted neuron is selected. However, the GMDH-NN only considers the two variable relations by ignoring the individual effect of each variable. The Architecture Group Method of Data Handling type Neural Network (RGMDH-NN), an improved form of GMDH-NN, is used which simulates not only the two-variable relation but also their individuals. The model for RGMDH-NN is described in the following equation:

$$y_t = a + \sum_{i=1}^{r} b_i x_i + \sum_{i=1}^{r} \sum_{j=1}^{r} c_{ij} x_i x_j \quad (10)$$

The rest of the procedure of RGMDH-NN is the same as GMDH-NN. The selected neuron with minimum Mean Square Error is transferred in the next layer by using the transfer functions listed in Table 1. The coefficients of each neuron are estimated with regularized least square estimation method as this method of estimation has the ability to solve the multicollinearity problem which is usually the inherited part of time series data with multiple lagged variables.

Radial basis function neural network: to predict the denoised IMFs, nonlinear neural network, i.e., Radial Basis

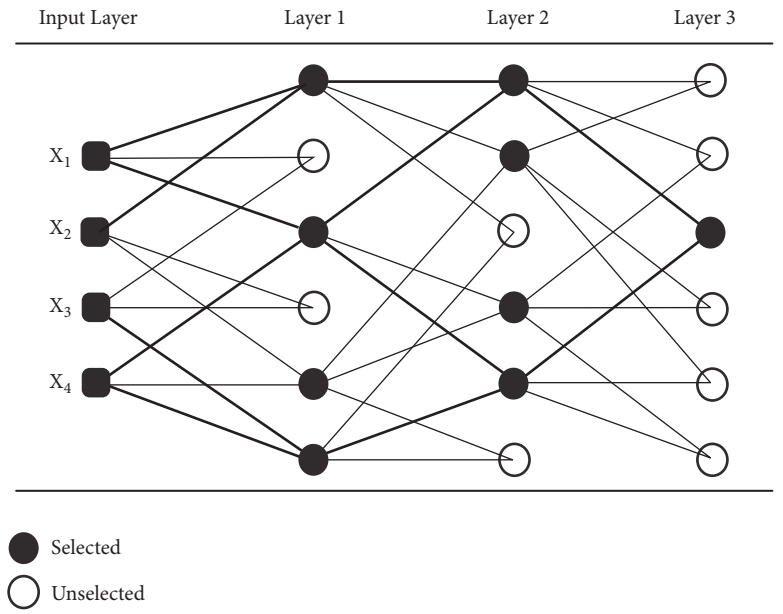


Figure 2: Architecture of GMDH-type neural network (NN) algorithms.

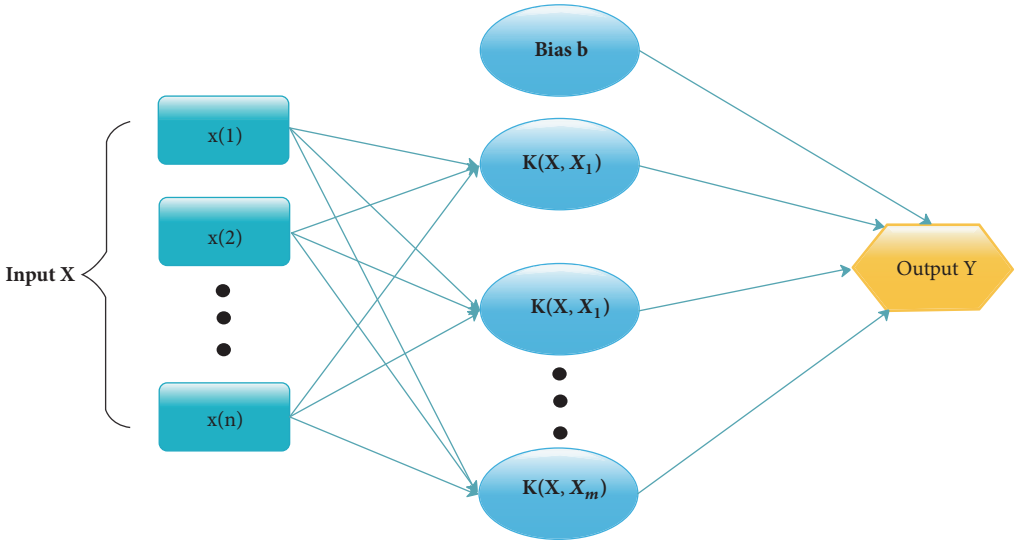


Figure 3: Topological structure of radial basis function.

Function (RBFNN), is also adopted. The reason for selecting RBFNN is that it has a nonlinear structure to find relation between lagged variables. The RBFNN is a three-layer feed-forward neural network which consists of an input layer, a hidden layer, and an output layer. The whole algorithm is illustrated in Figure 3. Unlike GMDH-NN, RBFNN takes all inputs in each neuron with corresponding weights and then hidden layer transfers the output by using radial basis function with weights to output. The sigmoid basis function is used to transfer the complex relation between lagged variables as follows:

$$\emptyset_i(x) = \frac{1}{1 + \exp\left(b^T x - b_0\right)} \tag{11}$$

3. Case Study and Experimental Design

Selection of Study Area. In this study, the Indus Basin System (IBS), known to be the largest river system in Pakistan, is considered which plays a vital role in the power generation and irrigation system. The major tributaries of this river are River Jhelum, River Chenab, and River Kabul. These rivers get their inflows mostly from rainfall, snow, and glacier melt. As in Pakistan, glaciers covered 13,680 km^2 area in which estimated 13% of the areas are covered by Upper Indus Basin (UIB) [33]. About 50% melted water from these 13% areas adds the significant contribution of water in these major rivers. The Indus river and its tributaries cause flooding due to glacier and snow melting and rainfall [34]. The major events

Figure 4: Rivers and irrigation network of Pakistan.

of flood usually occur in summer due to heavy monsoon rainfall which starts from July and end in September. It was reported [35] that, due to excessive monsoon rainfall in 2010, floods have been generated in IBS which affected 14 million people and around 20,000,000 inhabitants were displaced. Moreover, surface water system of Pakistan is also based on flows of IBS and its tributaries [36]. Pappas [37] mentioned that around 65% of agriculture land is irrigated with the Indus water system. Therefore, for effective water resources management and improving sustainable economic and social development and for proactive and integrated flood management, there is a need to appropriately analyze and predict the rivers inflow data of IBS and its tributaries.

Data. To thoroughly investigate the proposed models, four rivers' inflow data is used in this study which is comprised of daily rivers inflow (1^{st}-January to 19^{th}-June) for the period of 2015-2018. We consider the main river inflow of Indus at Tarbela with its two principal, one left and one right, bank tributaries [38]: Jhelum inflow at Mangla, Chenab at Marala, and Kabul at Nowshera, respectively (see Figure 4). Data is measured in 1000 cusecs. The rivers inflow data was acquired from the site of Pakistan Water and Power Development Authority (WAPDA).

Comparison of Proposed Study with Other Methods. The proposed models are compared with other prediction approaches by considering with and without principals of denoising and decomposition. For that purpose, the following types of models are selected:

(I) Without denoising and decomposing, only single statistical model is selected, i.e., ARIMA (for convenience, we call one-stage model 1-S) as used in [8].

(II) Only denoised based models: in this stage, the noise removal capabilities of WA and EMD are assessed. The wavelet based models are WA-ARIMA, WA-RBFNN, and WA-RGMDH whereas the empirical mode decomposition based models are EMD-ARIMA, EMD-RBFNN, and EMD-RGMDH. The different prediction models are chosen for the comparison of traditional statistical models with artificial intelligence based models as RBFN and RGMDH (for convenience, we call two-stage model 2-S). The 2-S selected models for comparison are used from [15, 17] for the comparison with the proposed model.

(III) With denoising and decomposition (existing method): for that purpose, three-stage EMD-EEMD-MM model is used from [2] for the comparison with proposed models. Under this, the multiple models are selected by keeping the prediction characteristics similar to proposed model for comparison purpose (for convenience, we call three-stage model 3-S).

Evaluation Criteria. The prediction accuracy of models is assessed using three evaluation measures such as Mean Relative Error (MRE), Mean Absolute Error (MAE), and Mean Square Error (MSE). The following are their equations, respectively:

$$MRE = \frac{1}{n}\frac{\sum_{t=1}^{n}\left|f(t) - \widehat{f(t)}\right|}{f(t)} \tag{12}$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n}\left|f(t) - \widehat{f(t)}\right| \tag{13}$$

and

$$MSE = \frac{1}{n}\sum_{t=1}^{n}\left(f(t) - \widehat{f(t)}\right)^2 \tag{14}$$

All proposed and selected models are evaluated using these criteria. Moreover, in GMDH-NN and RGMDH-NN models, neurons are selected according to MSE.

4. Results

D-stage results: the results of two noise removal filters, i.e., WA and EMD, are described below.

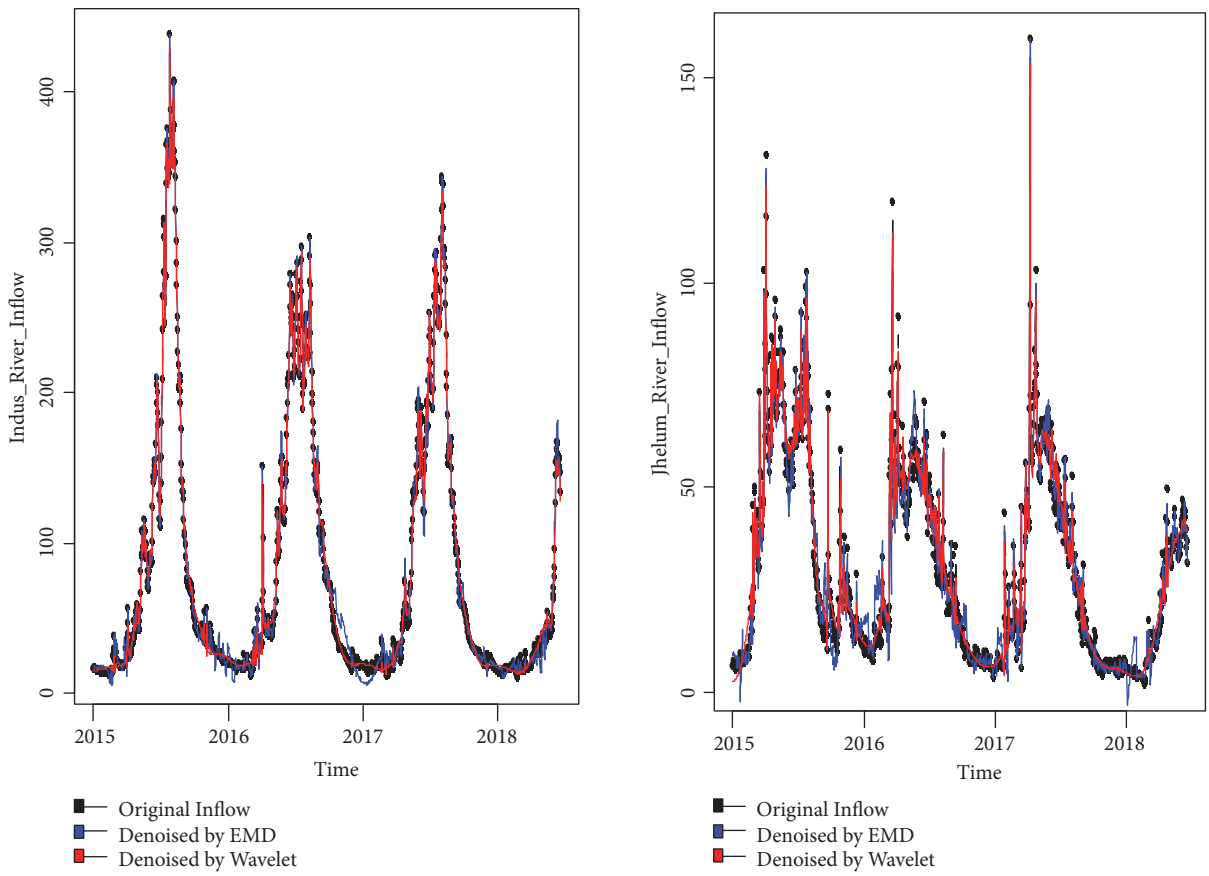


FIGURE 5: The denoised series of the two hydrological time series of Indus and Jhelum rivers inflow. The figure shows the denoised results obtained through the EMD-based threshold method (in red color) and the wavelet analysis-based threshold method (in blue color).

Wavelet based denoised: after calculating the approximations from (1) and details from (2), the hard and soft rule of thresholding are used to remove noises from hydrological time series coefficients. Both hard and soft rules are calculated from (3) and (4) respectively. On behalf of lower MSE, hard threshold based denoised series are reconstructed through (5) for WA.

EMD-based threshold: to remove noises through EMD, intrinsic mode functions are calculated from (7), and then hard and soft thresholds are used to denoise the calculated IMFs except the last two IMFs as, due to smooth and low frequency characteristics, there is no need to denoise the last two IMFs. Hard threshold based denoised IMFs are further used to reconstruct the noise free hydrological time series data from (8).

The WA and EMD based denoised Indus and Jhelum rivers inflow are shown in Figure 5. The statistical measures including mean ($\overline{x}$), standard deviation (σ), MRE, MAE, and MSE of original and denoised series for all case studies of both noise removal methods are presented in Table 2. The results show that the statistical measures are almost the same for both denoising methods except MSE, as for Indus and Jhelum inflow, WA-based denoised series have lower MSE than EMD; however, for Kabul and Chenab inflow, EMD-based denoised series have lower MSE than WA-based denoised series. Overall, it was concluded that both methods have equal performance in denoising the hydrological time series data. In decomposing stage, both of WA and EMD based denoised series are separately used as input to derive the time varying characteristics in terms of high and low frequencies.

Decompose-stage results: to extract the local time varying features from denoised hydrological data, the WA/EMD-based denoised hydrological time series data are further decomposed into nine IMFs and one residual. The CEEMDAN decomposition method is used to extract the IMFs from all four rivers. EMD-based denoised-CEEMDAN-based decomposed results of Indus and Jhelum rivers inflow are shown in Figure 6 whenever WA-CEEMDAN-based noise free decomposed results of Indus and Jhelum rivers inflow are shown in Figure 7. All four rivers are decomposed into nine IMFs and one residual showing similar characteristics for both methods. The extracted IMFs show the characteristics of hydrological time series data where the starting IMFs represent the higher frequency whereas last half IMFs show the low frequencies and residual are shown as trends as shown in Figures 6 and 7. The amplitude of white noise is set as 0.2 as in [2] and numbers of ensemble members are selected as maximum which is 1000.

Table 2: Statistical measures of WA- and EMD-based denoised rivers inflow of four hydrological time series data sets.

River Inflow	Mode	$\overline{x}$	σ	MRE	MAE	MSE
Indus Inflow	Original series	80.2194	87.5044			
	EMD	80.5931	87.3925	3.9118	0.1275	36.7636
	Wavelet	80.2267	86.1632	3.8188	0.0987	22.9626
Jhelum Inflow	Original series	30.2001	23.6743			
	EMD	30.1412	23.1641	2.7118	0.1666	16.4864
	Wavelet	30.2023	22.7799	2.5579	0.1418	10.8837
Kabul Inflow	Original series	29.1746	25.2352			
	EMD	25.23524	25.1181	2.5474	0.2036	12.5216
	Wavelet	29.18118	24.29148	2.7386	0.1615	12.2447
Chenab Inflow	Original series	31.9557	29.4916			
	EMD	32.0024	29.2734	2.271784	0.1470	10.6797
	Wavelet	31.9585	28.2591	3.1958	0.17228	17.8353

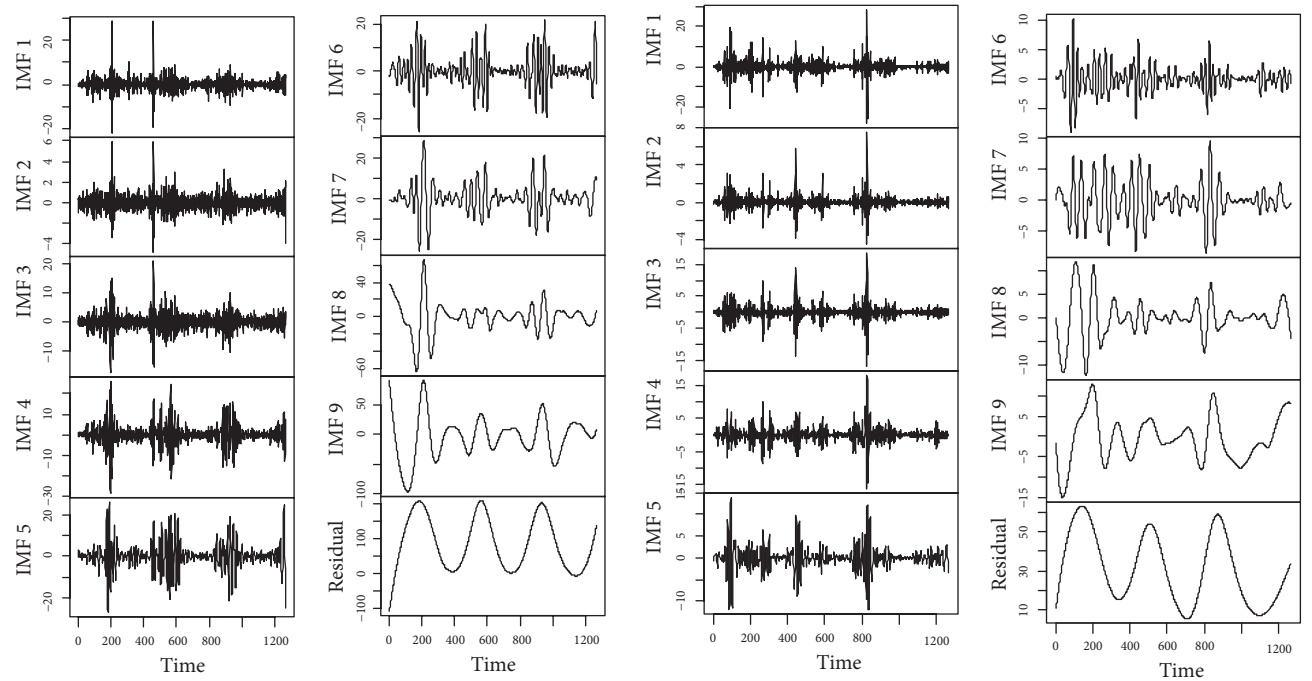


Figure 6: The EMD-CEEMDAN decomposition of Indus (left) and Jhelum rivers inflow (right). The two series are decomposed into nine IMFs and one residue.

P-step results: for all extracted IMFs and residual, three methods of predictions are adopted to get the more precise and near-to-reality results. For that reason, one traditional statistical method, i.e., ARIMA (p, d, q), with two other nonlinear ML methods, i.e., GMDH-NN and RBFNN, are used to predict the IMFs and residuals of all four river inflows. The rivers inflow data of all four rivers are split: 70% for training set and 30% for testing set. The parameters and structure of models are estimated using 886 observations of rivers inflow. The validity of proposed and selected models is tested using 30% data of rivers inflow. After successful estimation of multimodels on each IMF and residual, the best method with minimum MRE, MAE, and MSE is selected for each IMF prediction. The testing results of proposed models with comparison to all other models for all four rivers' inflow, i.e., Indus inflow, Jhelum inflow, Chenab inflow, and Kabul inflow, are presented in Table 3. The proposed EMD-CEEMDAN-MM and WA-CEEMDAN-MM model prediction results fully demonstrate the effectiveness for all 4 cases with minimum MRE, MAE, and MSE compared to all 1-S [8], 2-S [15, 17], and 3-S [2] evaluation models. However, overall, the proposed WA-CEEMDAN-MM model attains the lowest MSE as compared to other EMD-CEEMDAN-MM proposed models. The worst predicted model is 1-S, i.e., ARIMA, without denoising and without decomposing the hydrological time series data with highest MSE. The predicted graphs of proposed model, i.e., EMD-CEEMDAN-MM, with comparison to 2-S models, i.e., with EMD based denoised for

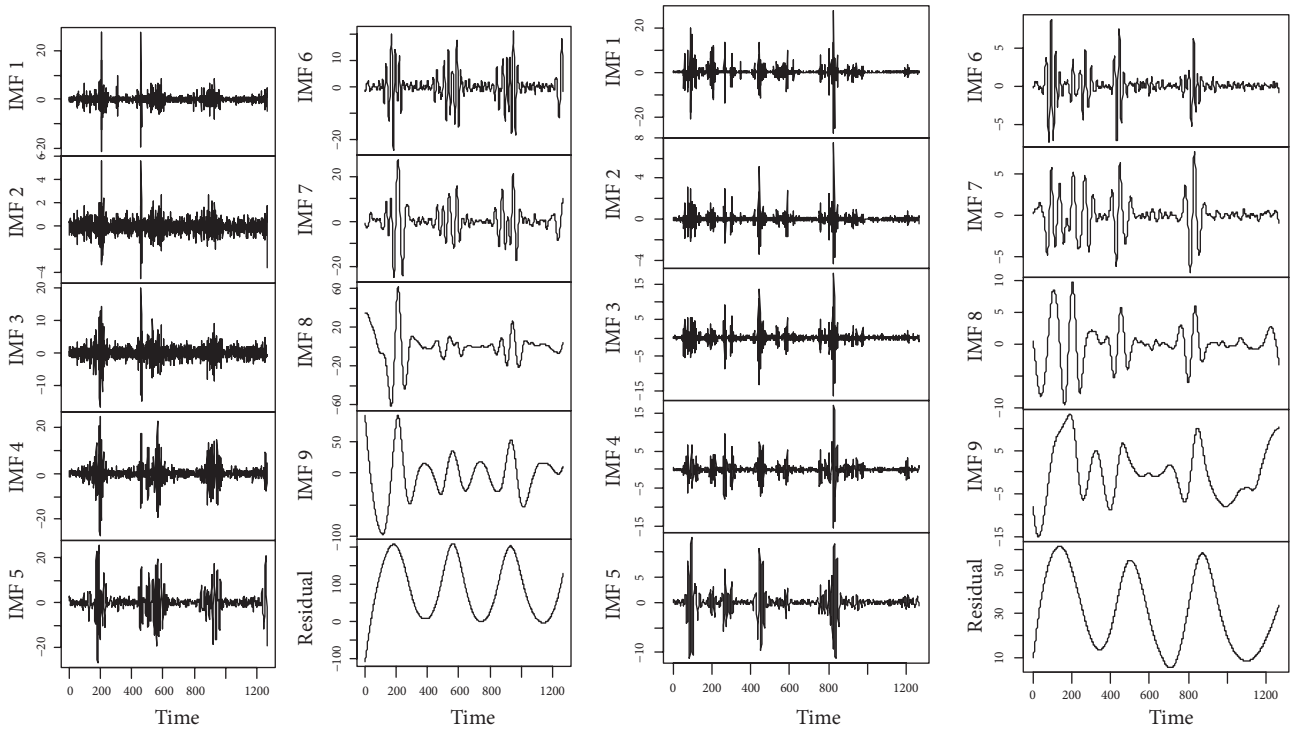


Figure 7: The WA-CEEMDAN decomposition of Indus (left) and Jhelum rivers inflow (right). The two series are decomposed into nine IMFs and one residue.

Indus and Jhelum river inflow, are shown in Figure 8 and WA-CEEMDAN-MM with comparison to 2-S models, i.e., with WA based denoised, are shown in Figure 9.

To improve the prediction accuracy of complex hydrological time series data from simple time series models one can take the advantage from three principals of "denoising," "decomposition," and "ensembling the predicted results." The 2-S model, with simple ARIMA and GMDH, can perform well as compared to 2-S models with complex models and 1-S models by optimal decomposition methods. Moreover, with addition to extracting time varying frequencies from denoised series, one can get the more precise results over 2-S models. However, from Table 3, it can be concluded that the proposed WA-CEEMDAN-MM and EMD-CEEMDAN-MM models perform more efficiently to predict the hydrological time series data by decreasing the complexity of hydrological time series data and enhancing the prediction performance over 1-S, 2-S, and 3-S existing models.

The following conclusions are drawn based on the testing error presented in Table 3.

Overall comparison: the overall performances of proposed models WA-CEEMDAN-MM and WA-CEEMDAN-MM are better than all other evaluation models selected from the study [2, 8, 15, 17] with the lowest MAE, MRE, and MSE values for all case studies. However, among two proposed models, WA-CEEMDAN-MM performs well by attaining on average 8.49%, 24.19%, and 5.43% lowest MAE, MRE, and MSE values, respectively, for all four rivers' inflow prediction as compared to EMD-CEEMDAN-MM as listed in Table 3. It is shown that both proposed models perform well with comparison to 1-S, 2-S, and existing 3-S. Moreover, it is also noticed that most of IMFs are precisely predicted with simple traditional statistical ARIMA model except the first two IMFs as the first IMFs presented high frequencies showing more volatile time varying characteristics with the rest of IMFs. However, overall WA-CEEMDAN-MM is more accurate in predicting the rivers inflow.

Comparison of proposed models with other only denoised series models: removing the noise through WA- and EMD-based threshold filters before statistical analysis improved the prediction accuracy of complex hydrological time series. It can be observed from Table 3 that the MAE, MRE, and MSE values of four cases perform well for 2-S model as compared to 1-S model in both WA and EMD based denoised inputs. However, like overall performance of WA-CEEMDAN-MM, WA-based denoised models perform well compared to EMD-based denoised. Moreover, with denoised series, the several statistical (simple) and machine learning (complex) methods are adopted to further explore the performances between simple and complex methods to predict inflows. This can be seen from Table 3, where WA-RBFN and EMD-RBFN perform the worst compared to WA-ARIMA, WA-RGMDH, EMD-ARIMA, and EMD-RGMDH. This means that with denoising the hydrological series one can move towards simple models as compared to complex models like radial basis function neural network. WA-RGMDH and EMD-RGMDH attain the highest accuracy among all 2-S models.

Comparison of proposed models with other denoised and decomposed models: in addition to denoising, the decomposition of hydrological time series strategy effectively enhances

Table 3: Evaluation index of testing prediction error of proposed models (EMD-CEEMDAN-MM and WA-CEEMDAN-MM) with all selected models for all four case studies.

River Inflow	Model Name	Models	MRE	MAE	MSE
Indus Inflow	1-S	ARIMA	4.2347	0.0685	64.7141
	2-S	WA-ARMA	3.2862	0.0430	53.4782
		WA-RGMDH	3.2548	0.0393	46.7382
		WA-RBFN	20.1949	0.2598	2301.772
		EMD-ARMA	4.9898	0.0960	76.1440
		EMD-RGMDH	4.9653	0.0915	76.0884
		EMD-RBFN	34.3741	0.7762	3931.601
	3-S	EMD-EEMD-MM	5.2710	0.1721	44.0115
		WA-CEEMDAN-MM	**1.5410**	**0.0349**	**5.5734**
		EMD-CEEMDAN-MM	**1.8009**	**0.0462**	**6.6983**
Jhelum Inflow	1-S	ARMA	3.5224	0.1201	47.5529
	2-S	WA-ARMA	2.6129	0.0748	37.1441
		WA-RGMDH	2.6208	0.0773	37.7954
		WA-RBFN	9.8608	0.7714	180.7443
		EMD-ARMA	3.7354	0.1551	48.3164
		EMD-RGMDH	3.7357	0.1620	48.3606
		EMD-RBFN	2.8822	0.2506	51.9916
	3-S	EMD-EEMD-MM	2.0096	0.1269	7.3565
		WA-CEEMDAN-MM	**1.1805**	**0.0457**	**6.8225**
		EMD-CEEMDAN-MM	**1.4480**	**0.0642**	**7.7709**
Kabul Inflow	1-S	ARMA	2.4910	0.0883	25.0136
	2-S	WA-ARMA	1.9999	0.0592	20.6874
		WA-RGMDH	2.0794	0.0729	21.0612
		WA-RBFN	1.6565	0.0997	13.3554
		EMD-ARMA	2.9538	0.1484	28.5767
		EMD-RGMDH	3.0114	0.2280	28.9351
		EMD-RBFN	4.9355	0.7613	69.9346
	3-S	EMD-EEMD-MM	1.8758	0.3166	5.8020
		WA-CEEMDAN-MM	**0.7664**	**0.0363**	**2.1072**
		EMD-CEEMDAN-MM	**0.9599**	**0.0861**	**2.7636**
Chenab Inflow	1-S	ARMA	5.4157	0.4646	108.185
	2-S	WA-ARMA	3.9652	0.1087	84.2359
		WA-RGMDH	3.6147	0.0943	81.6493
		WA-RBFN	4.1424	0.2757	47.6184
		EMD-ARMA	4.7971	0.2721	100.7013
		EMD-RGMDHA	4.4812	0.1865	95.6680
		EMD-RBFN	10.8228	2.1666	284.5627
	3-S	EMD-EEMD-MM	2.7172	0.2298	14.5191
		WA-CEEMDAN-MM	**1.6940**	**0.0705**	**13.5702**
		EMD-CEEMDAN-MM	**1.9345**	**0.1105**	**14.067**

the prediction accuracy by reducing the complexity of hydrological data in multiple dimensions. It is shown from Table 3 that the 3-S (existing) performs better as on average for all four rivers MAE, MRE, and MSE values are 13.76%, -6.55%, and 54.79%, respectively, lower than 1-S model and 63.40, 64.76%, and 96.78% lower than 2-S model (EMD-RBFNN). Further research work can be done to explore the ways to reduce the mathematical complexity of separate denoising and decomposition like only single filter which not only denoises but also decomposes the hydrological time series data with the same filter to effectively predict or simulate the data.

5. Conclusion

The accurate prediction of hydrological time series data is essential for water supply and water resources purposes. Considering the instability and complexity of hydrological

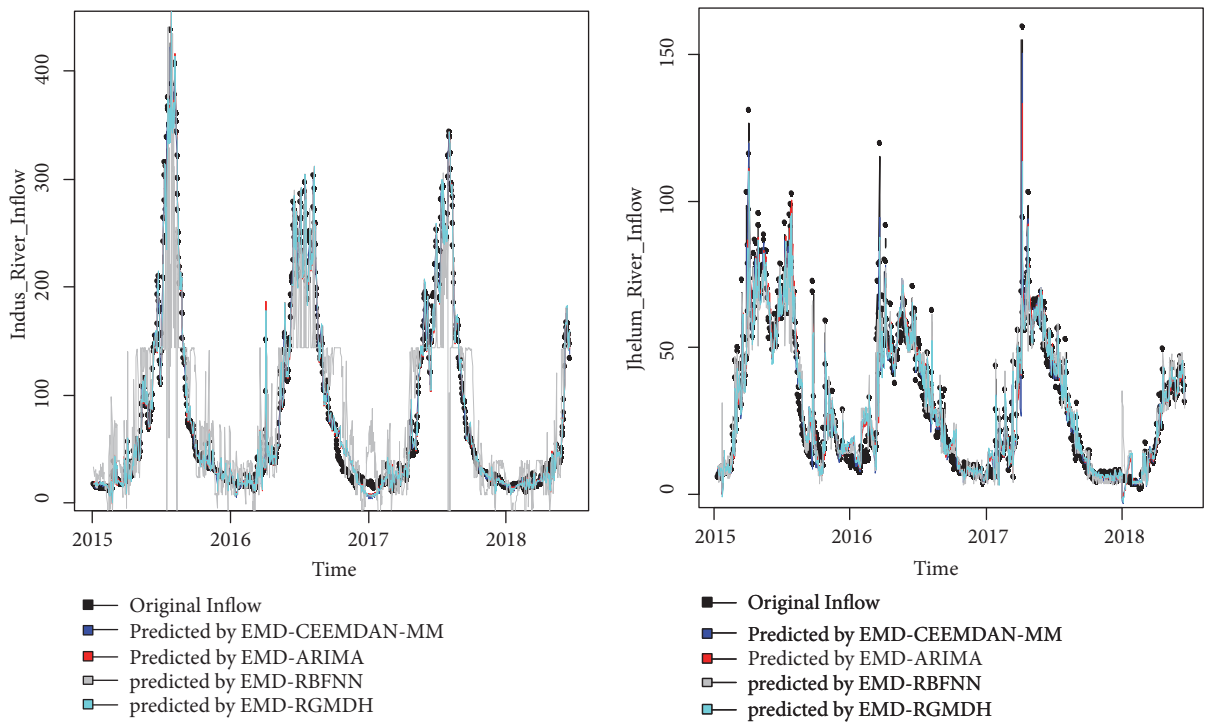


Figure 8: Prediction results of Indus and Jhelum rivers inflow using proposed EMD-CEEMDAN-MM with comparison to other EMD based denoised and predicted models.

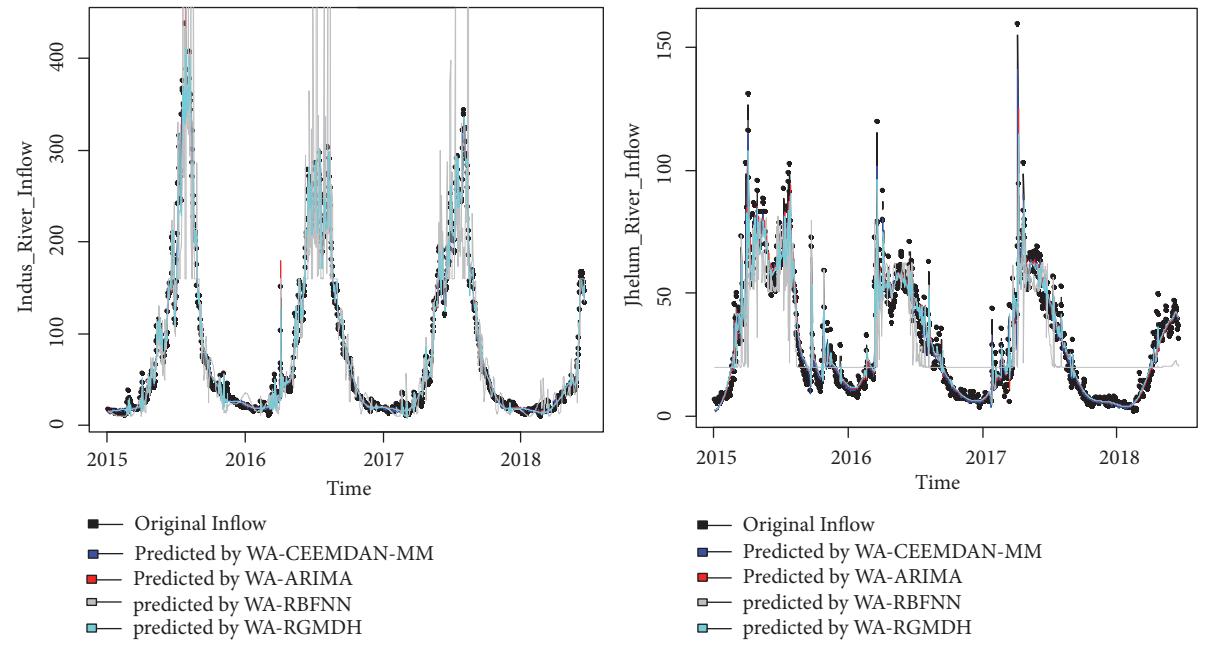


Figure 9: Prediction results of Indus and Jhelum river inflow using proposed WA-CEEMDAN-MM with comparison to WA based denoised predicted models.

time series, some data preprocessing methods are adopted with the aim to enhance the prediction of such stochastic data by decomposing the complexity of hydrological time series data in an effective way. This research proposed two new methods with three stages as "denoised," decomposition, and prediction and summation, named as WA-CEEMDAN-MM and EMD-CEEMDAN-MM, for efficiently predicting the hydrological time series. For the verification of proposed methods, four cases of rivers inflow data from Indus Basin System are utilized. The overall results show that the proposed

hybrid prediction model improves the prediction performance significantly and outperforms some other popular prediction methods. Our two proposed, three-stage hybrid models show improvement in prediction accuracy with minimum MRE, MAE, and MSE for all four rivers as compared to other existing one-stage [8] and two-stage [15, 17] and three-stage [2] models. In summary, the accuracy of prediction is improved by reducing the complexity of hydrological time series data by incorporating the denoising and decomposition. In addition, these new prediction models are also capable of solving other nonlinear prediction problems.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

Acknowledgments

The authors extend their appreciation to the Deanship of Scientific Research at King Saud University for funding this work through research group no. RG-1437-027.

References

[1] D. P. Solomatine and A. Ostfeld, "Data-driven modelling: some past experiences and new approaches," *Journal of Hydroinformatics*, vol. 10, no. 1, pp. 3–22, 2008.

[2] C. Di, X. Yang, and X. Wang, "A four-stage hybrid model for hydrological time series forecasting," *PLoS ONE*, vol. 9, no. 8, Article ID e104663, 2014.

[3] Z. Islam, *Literature Review on Physically Based Hydrological Modeling [Ph. D. thesis]*, pp. 1–45, 2011.

[4] A. R. Ghumman, Y. M. Ghazaw, A. R. Sohail, and K. Watanabe, "Runoff forecasting by artificial neural network and conventional model," *Alexandria Engineering Journal*, vol. 50, no. 4, pp. 345–350, 2011.

[5] S. Riad, J. Mania, L. Bouchaou, and Y. Najjar, "Rainfall-runoff model usingan artificial neural network approach," *Mathematical and Computer Modelling*, vol. 40, no. 7-8, pp. 839–846, 2004.

[6] T. Peng, J. Zhou, C. Zhang, and W. Fu, "Streamflow Forecasting Using Empirical Wavelet Transform and Artificial Neural Networks," *Water*, vol. 9, no. 6, p. 406, 2017.

[7] G. E. P. Box and G. M. Jenkins, *Time Series Analysis: Forecasting and Control*, Holden-Day, San Francisco, Calif, USA, 1970.

[8] B. N. S. Ghimire, "Application of ARIMA Model for River Discharges Analysis," *Journal of Nepal Physical Society*, vol. 4, no. 1, pp. 27–32, 2017.

[9] Ö. Kişi, "Streamflow forecasting using different artificial neural network algorithms," *Journal of Hydrologic Engineering*, vol. 12, no. 5, pp. 532–539, 2007.

[10] C. A. G. Santos and G. B. L. D. Silva, "Daily streamflow forecasting using a wavelet transform and artificial neural network hybrid models," *Hydrological Sciences Journal*, vol. 59, no. 2, pp. 312–324, 2014.

[11] C. Wu, K. Chau, and Y. Li, "Methods to improve neural network performance in daily flows prediction," *Journal of Hydrology*, vol. 372, no. 1-4, pp. 80–93, 2009.

[12] T. Partal, "Wavelet regression and wavelet neural network models for forecasting monthly streamflow," *Journal of Water and Climate Change*, vol. 8, no. 1, pp. 48–61, 2017.

[13] Z. M. Yaseen, M. Fu, C. Wang, W. H. Mohtar, R. C. Deo, and A. El-shafie, "Application of the Hybrid Artificial Neural Network Coupled with Rolling Mechanism and Grey Model Algorithms for Streamflow Forecasting Over Multiple Time Horizons," *Water Resources Management*, vol. 32, no. 5, pp. 1883–1899, 2018.

[14] M. Rezaie-Balf and O. Kisi, "New formulation for forecasting streamflow: evolutionary polynomial regression vs. extreme learning machine," *Hydrology Research*, vol. 49, no. 3, pp. 939–953, 2018.

[15] H. Liu, C. Chen, H.-Q. Tian, and Y.-F. Li, "A hybrid model for wind speed prediction using empirical mode decomposition and artificial neural networks," *Journal of Renewable Energy*, vol. 48, pp. 545–556, 2012.

[16] Z. Qu, K. Zhang, J. Wang, W. Zhang, and W. Leng, "A Hybrid model based on ensemble empirical mode decomposition and fruit fly optimization algorithm for wind speed forecasting," *Advances in Meteorology*, 2016.

[17] Y. Sang, "A Practical Guide to Discrete Wavelet Decomposition of Hydrologic Time Series," *Water Resources Management*, vol. 26, no. 11, pp. 3345–3365, 2012.

[18] N. E. Huang, Z. Shen, S. R. Long et al., "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 454, Article ID 1971, pp. 903–995, 1998.

[19] Z. H. Wu and N. E. Huang, "A study of the characteristics of white noise using the empirical mode decomposition method," *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 460, no. 2046, pp. 1597–1611, 2004.

[20] Z. Wang, J. Qiu, and F. Li, "Hybrid Models Combining EMD/EEMD and ARIMA for Long-Term Streamflow Forecasting," *Water*, vol. 10, no. 7, p. 853, 2018.

[21] N. A. Agana and A. Homaifar, "EMD-based predictive deep belief network for time series prediction: an application to drought forecasting," *Hydrology*, vol. 5, no. 1, p. 18, 2018.

[22] A. Kang, Q. Tan, X. Yuan, X. Lei, and Y. Yuan, "Short-term wind speed prediction using EEMD-LSSVM model," *Advances in Meteorology*, 2017.

[23] Z. H. Wu and N. E. Huang, "Ensemble empirical mode decomposition: a noise-assisted data analysis method," *Advances in Adaptive Data Analysis (AADA)*, vol. 1, no. 1, pp. 1–41, 2009.

[24] W.-C. Wang, K.-W. Chau, D.-M. Xu, and X.-Y. Chen, "Improving Forecasting Accuracy of Annual Runoff Time Series Using ARIMA Based on EEMD Decomposition," *Water Resources Management*, vol. 29, no. 8, pp. 2655–2675, 2015.

[25] H. Su, H. Li, Z. Chen, and Z. Wen, "An approach using ensemble empirical mode decomposition to remove noise from prototypical observations on dam safety," *SpringerPlus*, vol. 5, no. 1, 2016.

[26] X.-S. Jiang, L. Zhang, and M. X. Chen, "Short-term forecasting of high-speed rail demand: A hybrid approach combining ensemble empirical mode decomposition and gray support

vector machine with real-world applications in China," *Transportation Research Part C: Emerging Technologies*, vol. 44, pp. 110–127, 2014.

[27] S. Dai, D. Niu, and Y. Li, "Daily Peak Load Forecasting Based on Complete Ensemble Empirical Mode Decomposition with Adaptive Noise and Support Vector Machine Optimized by Modified Grey Wolf Optimization Algorithm," *Energies*, vol. 11, no. 1, p. 163, 2018.

[28] M. E. Torres, M. A. Colominas, G. Schlotthauer, and P. Flandrin, "A complete ensemble empirical mode decomposition with adaptive noise," in *Proceedings of the 36th IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 4144–4147, Prague, Czech Republic, May 2011.

[29] A. Jayawardena and A. Gurung, "Noise reduction and prediction of hydrometeorological time series: dynamical systems approach vs. stochastic approach," *Journal of Hydrology*, vol. 228, no. 3-4, pp. 242–264, 2000.

[30] M. Yang, Y. Sang, C. Liu, and Z. Wang, "Discussion on the Choice of Decomposition Level for Wavelet Based Hydrological Time Series Modeling," *Water*, vol. 8, no. 5, p. 197, 2016.

[31] J. Kim, C. Chun, and B. H. Cho, "Comparative analysis of the DWT-based denoising technique selection in noise-riding DCV of the Li-Ion battery pack," in *Proceedings of the 2015 9th International Conference on Power Electronics and ECCE Asia (ICPE 2015-ECCE Asia)*, pp. 2893–2897, Seoul, South Korea, June 2015.

[32] H. Ahmadi, M. Mottaghitalab, and N. Nariman-Zadeh, "Group Method of Data Handling-Type Neural Network Prediction of Broiler Performance Based on Dietary Metabolizable Energy, Methionine, and Lysine," *The Journal of Applied Poultry Research*, vol. 16, no. 4, pp. 494–501, 2007.

[33] A. S. Shakir and S. Ehsan, "Climate Change Impact on River Flows in Chitral Watershed," *Pakistan Journal of Engineering and Applied Sciences*, 2016.

[34] B. Khan, M. J. Iqbal, and M. A. Yosufzai, "Flood risk assessment of River Indus of Pakistan," *Arabian Journal of Geosciences*, vol. 4, no. 1-2, pp. 115–122, 2011.

[35] K. Gaurav, R. Sinha, and P. K. Panda, "The Indus flood of 2010 in Pakistan: a perspective analysis using remote sensing data," *Natural Hazards*, vol. 59, no. 3, pp. 1815–1826, 2011.

[36] A. Sarwar and A. S. Qureshi, "Water management in the indus basin in Pakistan: challenges and opportunities," *Mountain Research and Development*, vol. 31, no. 3, pp. 252–260, 2011.

[37] G. Pappas, "Pakistan and water: new pressures on global security and human health," *American Journal of Public Health*, vol. 101, no. 5, pp. 786–788, 2011.

[38] J. L. Wescoat, A. Siddiqi, and A. Muhammad, "Socio-Hydrology of Channel Flows in Complex River Basins: Rivers, Canals, and Distributaries in Punjab, Pakistan," *Water Resources Research*, vol. 54, no. 1, pp. 464–479, 2018.

Hindawi
Complexity
Volume 2019, Article ID 7408725, 16 pages
https://doi.org/10.1155/2019/7408725

Research Article

End-Point Static Control of Basic Oxygen Furnace (BOF) Steelmaking Based on Wavelet Transform Weighted Twin Support Vector Regression

Chuang Gao,[1,2] Minggang Shen ,[1] Xiaoping Liu,[3] Lidong Wang,[2] and Maoxiang Chu[2]

[1]*School of Materials and Metallurgy, University of Science and Technology Liaoning, Anshan, Liaoning, China*
[2]*School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan, Liaoning, China*
[3]*School of Information and Electrical Engineering, Shandong Jianzhu University, Jinan, Shandong, China*

Correspondence should be addressed to Minggang Shen; lnassmg@163.com

Received 6 July 2018; Revised 22 October 2018; Accepted 10 December 2018; Published 1 January 2019

Academic Editor: Michele Scarpiniti

A static control model is proposed based on wavelet transform weighted twin support vector regression (WTWTSVR). Firstly, new weighted matrix and coefficient vector are added into the objective functions of twin support vector regression (TSVR) to improve the performance of the algorithm. The performance test confirms the effectiveness of WTWTSVR. Secondly, the static control model is established based on WTWTSVR and 220 samples in real plant, which consists of prediction models, control models, regulating units, controller, and BOF. Finally, the results of proposed prediction models show that the prediction error bound with 0.005% in carbon content and 10°C in temperature can achieve a hit rate of 92% and 96%, respectively. In addition, the double hit rate of 90% is the best result by comparing with four existing methods. The results of the proposed static control model indicate that the control error bound with 800 Nm^3 in the oxygen blowing volume and 5.5 *tons* in the weight of auxiliary materials can achieve a hit rate of 90% and 88%, respectively. Therefore, the proposed model can provide a significant reference for real BOF applications, and also it can be extended to the prediction and control of other industry applications.

1. Introduction

With the development of end-point control technology for basic oxygen furnace (BOF), the static control model can be established to overcome the randomness and inconsistency of the artificial experience control models. According to the initial conditions of hot metal, the relative calculations can be carried out to guide production. The end-point hit rate would be improved through this approach. Unfortunately, the control parameters would not be adjusted during the smelting process, which restricts the further improvement of the end-point hit rate. To solve this problem, the sublance based dynamic control model could be adopted by using the sublance technology. By combining the static model with the dynamic model, the end-point hit rate of BOF could be guaranteed. The establishment of the static control model is the foundation of the dynamic model. The accuracy of the static model will directly affect the hit rates of the dynamic control model, thus it plays an important role in the parameter optimization of BOF control process. Therefore, the static control model is still a practical and reliable technology to guide the production and improve the technology and management level of steelmaking plants.

In recent years, some significant developments of BOF prediction and control modelling have been achieved. Blanco et al. [1] designed a mixed controller for carbon and silicon in a steel converter in 1993. In 2002, three back propagation models are adopted to predict the end-blow oxygen volume and the weight of coolant additions [2]. In 2006, a dynamic model is constructed to predict the carbon content and temperature for the end-blow stage of BOF [3]. Based on multivariate data analysis, the slopping prediction was proposed by Brämming et al. [4]. In 2014, the multi-level recursive regression model was established for the prediction of end-point phosphorus content during BOF steelmaking process [5]. An antijamming endpoint prediction model

of extreme learning machine (ELM) was proposed with evolving membrane algorithm [6]. By applying input variables selection technique, the input weighted support vector machine modelling was proposed [7], and then the prediction model was established on the basis of improving a case-based reasoning method [8]. The neural network prediction modellings [9–12] were carried out to achieve aimed end-point conditions in liquid steel. A fuzzy logic control scheme was given for the basic oxygen furnace in [13]. Most of these achievements are based on the statistical and intelligent methods.

As an intelligent method, Jayadeva et. al. [14] proposed a twin support vector machine (TSVM) algorithm in 2007. The advantage of this method is that the computational complexity of modelling can be reduced by solving two quadratic programming problems instead of one in traditional method. It is also widely applied to the classification applications. In 2010, Peng [15] proposed a twin support vector regression (TSVR), which can be used to establish the prediction model for industrial data. After that, some improved TSVR methods [16–22] were proposed. By introducing a K-nearest neighbor (KNN) weighted matrix into the optimization problem in TSVR, the modified algorithms [16, 19] were proposed to improve the performance of TSVR. A v-TSVR [17] and asymmetric v-TSVR [20] were proposed to enhance the generalization ability by tuning new model parameters. To solve the ill-conditioned problem in the dual objective functions of the traditional TSVR, an implicit Lagrangian formulation for TSVR [18] was proposed to ensure that the matrices in the formulation are always positive semidefinite matrices. Parastalooi et al. [21] added a new term into the objective function to obtain structural information of the input data. By comparing with the neural network technology, the disadvantage of neural network is that the optimization process may fall into the local optimum. The optimization of TSVR is a pair of quadratic programming problems (QPPs), which means there must be a global optimal solution for each QPP. All above modified TSVR algorithms are focused on the improvements of the algorithm accuracy and the computation speed. Currently, the TSVR algorithm has never been adopted in the BOF applications. Motivated by this, the TSVR algorithm can be used to establish a BOF model.

Wavelet transform can fully highlight the characteristics of some aspects of the problem, which has attracted more and more attention and been applied to many engineering fields. In this paper, the wavelet transform technique is used to denoise the output samples during the learning process, and it is a new application of the combination of wavelet transform and support vector machine method. Then, a novel static control model is proposed based on wavelet transform weighted twin support vector regression (WTWTSVR), which is an extended model of our previous work. In [23], we proposed an end-point prediction model with WTWTSVR for the carbon content and temperature of BOF, and the accuracy of the prediction model is expected. However, the prediction model cannot be used to guide real BOF production directly. Hence, a static control model should be established based on the prediction model to calculate the oxygen volume and the weight of auxiliary raw materials, and the accuracy of the calculations affects the quality of the steel. Therefore, the proposed control model can provide a guiding significance for real BOF production. It is also helpful to other metallurgical prediction and control applications. To improve the performance of the control model, an improvement of the traditional TSVR algorithm is carried out. A new weighted matrix and a coefficient vector are added into the objective function of TSVR. Also, the parameter ε in TSVR is not adjustable anymore, which means it is a parameter to be optimized. Finally, the static control model is established based on the real datasets collected from the plant. The performance of the proposed method is verified by comparing with other four existing regression methods. The contributions of this work include the following. (1) It is the first attempt to establish the static control model of BOF by using the proposed WTWTSVR algorithm. (2) New weighted matrix and coefficient vector are determined by the wavelet transform theory, which gives a new idea for the optimization problem in TSVR areas. (3) The proposed algorithm is an extension of the TSVR algorithm, which is more flexible and accurate to establish a prediction and control model. (4) The proposed control model provides a new approach for the applications of BOF control. The application range of the proposed method could be extended to other metallurgical industries such as the prediction and control in the blast furnace process and continuous casting process.

Remark 1. The main difference between primal KNNWTSVR [16] and the proposed method is that the proposed algorithm utilizes the wavelet weighted matrix instead of KNN weighted matrix for the squared Euclidean distances from the estimated function to the training points. Also, a wavelet weighted vector is introduced into the objective functions for the slack vectors. ε_1 and ε_2 are taken as the optimized parameters in the proposed algorithm to enhance the generalization ability. Another difference is that the optimization problems of the proposed algorithm are solved in the Lagrangian dual space and that of KNNWTSVR are solved in the primal space via unconstrained convex minimization.

Remark 2. By comparing with available weighted technique like K-nearest neighbor, the advantages of the wavelet transform weighting scheme are embodied in the following two aspects:

(1) The Adaptability of the Samples. The proposed method is suitable for dealing with time/spatial sequence samples (such as the samples adopted in this paper) due to the character of wavelet transform. The wavelet transform inherits and develops the idea of short-time Fourier transform. The weights of sample points used in the proposed algorithm are determined by calculating the difference between the sample values and the wavelet-regression values for Gaussian function, which can mitigate the noise, especially the influence of outliers. While KNN algorithm determines the weight of the sample points by calculating the number of adjacent points (determined by Euclidian distance), which is more suitable for the samples of multi-points clustering type distribution.

(2) The Computational Complexity. KNN algorithm requires a large amount of computations, because the distance between each sample to all known samples must be computed to obtain its K nearest neighbors. By comparing with KNN weighting scheme, the wavelet transform weighting scheme has less computational complexity, because it is dealing with one dimensional output samples, and the computation complexity is proportional to the number of samples l. KNN scheme is dealing with the input samples, the computation complexity of KNN scheme is proportional to l^2. With the increasing of dimensions and number of the samples, it will have a large amount of computations.

Therefore, the wavelet transform weighting scheme is more competitive than KNN weighting scheme for the time sequence samples due to its low computational complexity.

2. Background

2.1. Description of BOF Steelmaking. BOF is used to produce the steel with wide range of carbon, alloy, and special alloy steels. Normally, the capacity of BOF is between 100 *tons* and 400 *tons*. When the molten iron is delivered to the converter through the rail, the desulphurization process is firstly required. In BOF, the hot metal and scrap, lime, and other fluxes are poured into the converter. The oxidation reaction is carried out with carbon, silicon, phosphorus, manganese, and some iron by blowing in a certain volume of oxygen. The ultimate goal of steelmaking is to produce the steel with specific chemical composition at suitable tapping temperature. The control of BOF is difficult because the whole smelting process is only half an hour, and there is no opportunity for sampling and analysis in the smelting process. The proportion of the iron and scrap is about 3:1 in BOF. The crane loads the waste into the container and then pours the molten iron into the converter. The water cooled oxygen lance enters the converter, the high purity oxygen is blown into BOF at 16000 cubic feet per minute, and the oxygen is reacted with carbon and other elements to reduce the impurity in the molten metal and converts it into a clean, high quality liquid steel. The molten steel is poured into the ladle and sent to the metallurgical equipment of the ladle [13].

Through the oxidation reaction of oxygen blown in BOF, the molten pig iron and the scrap can be converted into steel. It is a widely used steelmaking method with its higher productivity and low production cost [3]. However, the physical and chemical process of BOF is very complicated. Also, there are various types of steel produced in the same BOF, which means that the grade of steel is changed frequently. Therefore, the BOF modelling is a challenge task. The main objective of the modelling is to obtain the prescribed end-point carbon content and temperature.

2.2. Nonlinear Twin Support Vector Regression. Support vector regression (SVR) was proposed for the applications of regression problems. For the nonlinear case, it is based on the structural risk minimization and Vapnik ε-insensitive loss function. In order to improve the training speed of SVR, Peng proposed a TSVR algorithm in 2010. The difference between TSVR and SVR is that SVR solves one large QPP problem and TSVR solves two small QPP problems to improve the learning efficiency. Assume that a sample is an n-dimensional vector and the number of the samples is l, which can be expressed as $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l)$. Let $\mathbf{A} = [\mathbf{x}_1, \ldots, \mathbf{x}_l]^T \in R^{l\times n}$ be the input data set of training samples, $\mathbf{y} = [y_1, \ldots, \mathrm{y}_l]^T \in R^l$ be the corresponding output, and $\mathbf{e} = [1, \ldots, 1]^T$ be the ones vector with appropriate dimensions. Assume $K(.,.)$ denotes a nonlinear kernel function. Let $K(\mathbf{A}, \mathbf{A}^T)$ be the kernel matrix with order l and its (i, j)-th element $(i, j = 1, 2, \ldots, l)$ be defined by

$$\left[K\left(\mathbf{A}, \mathbf{A}^T\right)\right]_{i,j} = K\left(\mathbf{x}_i, \mathbf{x}_j\right) = \left(\Phi\left(\mathbf{x}_i\right) \cdot \Phi\left(\mathbf{x}_j\right)\right) \in R. \quad (1)$$

Here, the kernel function $K(\mathbf{x}, \mathbf{x}_i)$ represents the inner product of the nonlinear mapping functions $\Phi(\mathbf{x}_i)$ and $\Phi(\mathbf{x}_j)$ in the high dimensional feature space. Because there are various kernel functions, the performance comparisons of kernel functions will be discussed later. In this paper, the radial basis kernel function (RBF) is chosen as follows:

$$K\left(\mathbf{x}, \mathbf{x}_i\right) = \exp\left(-\frac{\left\|\mathbf{x} - \mathbf{x}_i\right\|^2}{2\sigma^2}\right), \quad (2)$$

where σ is the width of the kernel function. Let $K(\mathbf{x}^T, \mathbf{A}^T) = (K(\mathbf{x}, \mathbf{x}_1), K(\mathbf{x}, \mathbf{x}_2), \ldots, K(\mathbf{x}, \mathbf{x}_l))$ be a row vector in R^l. Then, two ε-insensitive bound functions $f_1(\mathbf{x}) = K(\mathbf{x}^T, \mathbf{A}^T)\boldsymbol{\omega}_1 + b_1$ and $f_2(\mathbf{x}) = K(\mathbf{x}^T, \mathbf{A}^T)\boldsymbol{\omega}_2 + b_2$ can be obtained, where $\boldsymbol{\omega}_1, \boldsymbol{\omega}_2 \in R^l$ are the normal vectors and $b_1, b_2 \in R$ are the bias values. Therefore, the final regression function $f(\mathbf{x})$ is determined by the mean of $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$, that is,

$$\begin{aligned} f\left(\mathbf{x}\right) &= \frac{1}{2}\left(f_1\left(\mathbf{x}\right) + f_2\left(\mathbf{x}\right)\right) \\ &= \frac{1}{2}K\left(\mathbf{x}^T, \mathbf{A}^T\right)\left(\boldsymbol{\omega}_1 + \boldsymbol{\omega}_2\right) + \frac{1}{2}\left(b_1 + b_2\right). \end{aligned} \quad (3)$$

Nonlinear TSVR can be obtained by solving two QPPs as follows:

$$\begin{aligned} \min_{\boldsymbol{\omega}_1, b_1, \boldsymbol{\xi}} \quad & \frac{1}{2}\left\|\mathbf{y} - \varepsilon_1\mathbf{e} - \left(K\left(\mathbf{A}, \mathbf{A}^T\right)\boldsymbol{\omega}_1 + b_1\mathbf{e}\right)\right\|^2 + c_1\mathbf{e}^T\boldsymbol{\xi} \\ \text{s.t.:} \quad & \mathbf{y} - \left(K\left(\mathbf{A}, \mathbf{A}^T\right)\boldsymbol{\omega}_1 + b_1\mathbf{e}\right) \geq \varepsilon_1\mathbf{e} - \boldsymbol{\xi}, \quad \boldsymbol{\xi} \geq 0e, \end{aligned} \quad (4)$$

and

$$\begin{aligned} \min_{\boldsymbol{\omega}_2, b_2, \boldsymbol{\gamma}} \quad & \frac{1}{2}\left\|\mathbf{y} + \varepsilon_2\mathbf{e} - \left(K\left(\mathbf{A}, \mathbf{A}^T\right)\boldsymbol{\omega}_2 + b_2\mathbf{e}\right)\right\|^2 + c_2\mathbf{e}^T\boldsymbol{\gamma} \\ \text{s.t.:} \quad & K\left(\mathbf{A}, \mathbf{A}^T\right)\boldsymbol{\omega}_2 + b_2\mathbf{e} - \mathbf{y} \geq \varepsilon_2\mathbf{e} - \boldsymbol{\gamma}, \quad \boldsymbol{\gamma} \geq 0\mathbf{e}, \end{aligned} \quad (5)$$

By introducing the Lagrangian function and Karush-Kuhn-Tucker conditions, the dual formulations of (4) and (5) can be derived as follows:

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & -\frac{1}{2}\boldsymbol{\alpha}^T\mathbf{G}\left(\mathbf{G}^T\mathbf{G}\right)^{-1}\mathbf{G}^T\boldsymbol{\alpha} + \mathbf{g}^T\mathbf{G}\left(\mathbf{G}^T\mathbf{G}\right)^{-1}\mathbf{G}^T\boldsymbol{\alpha} \\ & - \mathbf{g}^T\boldsymbol{\alpha} \\ \text{s.t.:} \quad & 0\mathbf{e} \leq \boldsymbol{\alpha} \leq c_1\mathbf{e}, \end{aligned} \quad (6)$$

and

$$\begin{aligned}\max_{\boldsymbol{\beta}} \quad & -\frac{1}{2}\boldsymbol{\beta}^T\mathbf{G}\left(\mathbf{G}^T\mathbf{G}\right)^{-1}\mathbf{G}^T\boldsymbol{\beta} - \mathbf{h}^T\mathbf{G}\left(\mathbf{G}^T\mathbf{G}\right)^{-1}\mathbf{G}^T\boldsymbol{\beta} \\ & + \mathbf{h}^T\boldsymbol{\beta} \\ \text{s.t.:} \quad & 0\mathbf{e} \le \boldsymbol{\beta} \le c_2\mathbf{e},\end{aligned} \tag{7}$$

where $\mathbf{G} = [K(\mathbf{A}, \mathbf{A}^T), \mathbf{e}]$, $\mathbf{g} = \mathbf{y} - \varepsilon_1\mathbf{e}$, and $\mathbf{h} = \mathbf{y} + \varepsilon_2\mathbf{e}$.

To solve the above QPPs (6) and (7), the vectors $\boldsymbol{\omega}_1, b_1$ and $\boldsymbol{\omega}_2, b_2$ can be obtained:

$$\begin{bmatrix}\boldsymbol{\omega}_1 \\ b_1\end{bmatrix} = \left(\mathbf{G}^T\mathbf{G}\right)^{-1}\mathbf{G}^T\left(\mathbf{g} - \boldsymbol{\alpha}\right), \tag{8}$$

and

$$\begin{bmatrix}\boldsymbol{\omega}_2 \\ b_2\end{bmatrix} = \left(\mathbf{G}^T\mathbf{G}\right)^{-1}\mathbf{G}^T\left(\mathbf{h} + \boldsymbol{\beta}\right). \tag{9}$$

By substituting the above results into (3), the final regression function can be obtained.

2.3. Nonlinear Wavelet Transform Based Weighted Twin Support Vector Regression

2.3.1. Model Description of Nonlinear WTWTSVR. In 2017, Xu el. al. [20] proposed the asymmetric ν-TSVR algorithm based on pinball loss functions. This new algorithm can enhance the generalization ability by tuning new model parameters. The QPPs of nonlinear asymmetric ν-TSVR were proposed as follows:

$$\begin{aligned}\min_{\boldsymbol{\omega}_1, b_1, \boldsymbol{\xi}, \varepsilon_1} \quad & \frac{1}{2}\left\|\mathbf{y} - \left(K\left(\mathbf{A}, \mathbf{A}^T\right)\boldsymbol{\omega}_1 + b_1\mathbf{e}\right)\right\|^2 + c_1\nu_1\varepsilon_1 \\ & + \frac{1}{l}c_1\mathbf{e}^T\boldsymbol{\xi} \\ \text{s.t.:} \quad & \mathbf{y} - \left(K\left(\mathbf{A}, \mathbf{A}^T\right)\boldsymbol{\omega}_1 + b_1\mathbf{e}\right) \\ & \ge -\varepsilon_1\mathbf{e} - 2\left(1 - p\right)\boldsymbol{\xi}, \quad \boldsymbol{\xi} \ge 0\mathbf{e}, \ \varepsilon_1 \ge 0\end{aligned} \tag{10}$$

and

$$\begin{aligned}\min_{\boldsymbol{\omega}_2, b_2, \varepsilon_2, \boldsymbol{\gamma}} \quad & \frac{1}{2}\left\|\mathbf{y} - \left(K\left(\mathbf{A}, \mathbf{A}^T\right)\boldsymbol{\omega}_2 + b_2\mathbf{e}\right)\right\|^2 + c_2\nu_2\varepsilon_2 \\ & + \frac{1}{l}c_2\mathbf{e}^T\boldsymbol{\xi}^* \\ \text{s.t.:} \quad & K\left(\mathbf{A}, \mathbf{A}^T\right)\boldsymbol{\omega}_2 + b_2\mathbf{e} - \mathbf{y} \ge -\varepsilon_2\mathbf{e} - 2p\boldsymbol{\xi}^*, \\ & \boldsymbol{\xi}^* \ge 0\mathbf{e}, \ \varepsilon_2 \ge 0\end{aligned} \tag{11}$$

where $c_1, c_2, \nu_1, \nu_2 \ge 0$ are regulating parameters and the parameter $p \in (0, 1)$ is used to apply a slightly different penalty for the outliers. The contributions of this method are that ε_1 and ε_2 are introduced into the objective functions of QPPs, and the parameters ν_1 and ν_2 are used to regulate the width of ε_1 and ε_2 tubes. Also, the parameter p is designed to give an unbalance weight for the slack vectors $\boldsymbol{\xi}$ and $\boldsymbol{\xi}^*$. Based on the concept of asymmetric ν-TSVR, a nonlinear wavelets transform based weighted TSVR is firstly proposed in this paper. By comparing with the traditional TSVR algorithm, the proposed algorithm introduces a wavelet transform based weighted matrix $\mathbf{D}_1$ and a coefficient vector $\mathbf{D}_2$ into the objective function of TSVR. Also, the parameters ε_1 and ε_2 are not adjustable by user anymore, which means they are both introduced into the objective functions. Simultaneously, the regularization is also considered to obtain the optimal solutions. Therefore, the QPPs of nonlinear WTWTSVR are proposed as follows:

$$\begin{aligned}\min_{\boldsymbol{\omega}_1, b_1, \boldsymbol{\xi}, \varepsilon_1} \quad & \frac{1}{2}\left(\mathbf{y} - \left(K\left(\mathbf{A}, \mathbf{A}^T\right)\boldsymbol{\omega}_1 + b_1\mathbf{e}\right)\right)^T\mathbf{D}_1\left(\mathbf{y} - \left(K\left(\mathbf{A}, \mathbf{A}^T\right)\boldsymbol{\omega}_1 + b_1\mathbf{e}\right)\right) + \frac{c_1}{2}\left(\boldsymbol{\omega}_1^T\boldsymbol{\omega}_1 + b_1^2\right) + c_2\left(\mathbf{D}_2^T\boldsymbol{\xi} + \nu_1\varepsilon_1\right) \\ \text{s.t.:} \quad & \mathbf{y} - \left(K\left(\mathbf{A}, \mathbf{A}^T\right)\boldsymbol{\omega}_1 + b_1\mathbf{e}\right) \ge -\varepsilon_1\mathbf{e} - \boldsymbol{\xi}, \quad \boldsymbol{\xi} \ge 0\mathbf{e}, \ \varepsilon_1 \ge 0,\end{aligned} \tag{12}$$

and

$$\begin{aligned}\min_{\boldsymbol{\omega}_2, b_2, \boldsymbol{\xi}^*, \varepsilon_2} \quad & \frac{1}{2}\left(\mathbf{y} - \left(K\left(\mathbf{A}, \mathbf{A}^T\right)\boldsymbol{\omega}_2 + b_2\mathbf{e}\right)\right)^T\mathbf{D}_1\left(\mathbf{y} - \left(K\left(\mathbf{A}, \mathbf{A}^T\right)\boldsymbol{\omega}_2 + b_2\mathbf{e}\right)\right) + \frac{c_3}{2}\left(\boldsymbol{\omega}_2^T\boldsymbol{\omega}_2 + b_1^2\right) + c_4\left(\mathbf{D}_2^T\boldsymbol{\xi}^* + \nu_2\varepsilon_2\right) \\ \text{s.t.:} \quad & K\left(\mathbf{A}, \mathbf{A}^T\right)\boldsymbol{\omega}_2 + b_2\mathbf{e} - \mathbf{y} \ge -\varepsilon_2\mathbf{e} - \boldsymbol{\xi}^*, \quad \boldsymbol{\xi}^* \ge 0\mathbf{e}, \ \varepsilon_2 \ge 0,\end{aligned} \tag{13}$$

where $c_1, c_2, c_3, c_4, \nu_1, \nu_2 \ge 0$ are regulating parameters, $\mathbf{D}_1$ is a diagonal matrix with the order of $l \times l$, and $\mathbf{D}_2$ is a coefficient vector with the length of $l \times 1$. The determination of $\mathbf{D}_1$ and $\mathbf{D}_2$ will be discussed later.

The first term in the objective function of (12) or (13) is used to minimize the sums of squared Euclidean distances from the estimated function $f_1(\mathbf{x}) = K(\mathbf{x}^T, \mathbf{A}^T)\boldsymbol{\omega}_1 + b_1$ or $f_2(\mathbf{x}) = K(\mathbf{x}^T, \mathbf{A}^T)\boldsymbol{\omega}_2 + b_2$ to the training points. The matrix $\mathbf{D}_1$ gives different weights for each Euclidean distance. The second term is the regularization term to avoid the overfitting problem. The third term minimizes the slack vector $\boldsymbol{\xi}$ or $\boldsymbol{\xi}^*$ and the width of ε_1-tube or ε_2-tube. The coefficient vector $\mathbf{D}_2$ is a penalty vector for the slack vector. To solve the problem in (12), the Lagrangian function can be introduced, that is,

$$\begin{aligned} &L(\boldsymbol{\omega}_1, b_1, \boldsymbol{\xi}, \varepsilon_1, \boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma) \\ &= \frac{1}{2}\left(\mathbf{y} - \left(K\left(\mathbf{A}, \mathbf{A}^T\right)\boldsymbol{\omega}_1 + b_1\mathbf{e}\right)\right)^T \\ &\cdot \mathbf{D}_1\left(\mathbf{y} - \left(K\left(\mathbf{A}, \mathbf{A}^T\right)\boldsymbol{\omega}_1 + b_1\mathbf{e}\right)\right) \\ &+ \frac{c_1}{2}\left(\boldsymbol{\omega}_1^T\boldsymbol{\omega}_1 + b_1^2\right) + c_2\left(\mathbf{D}_2^T\boldsymbol{\xi} + \nu_1\varepsilon_1\right) \\ &- \boldsymbol{\alpha}^T\left(\mathbf{y} - \left(K\left(\mathbf{A}, \mathbf{A}^T\right)\boldsymbol{\omega}_1 + b_1\mathbf{e}\right) + \varepsilon_1\mathbf{e} + \boldsymbol{\xi}\right) - \boldsymbol{\beta}^T\boldsymbol{\xi} \\ &- \gamma\varepsilon_1, \end{aligned} \tag{14}$$

where $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and γ are the positive Lagrangian multipliers. By differentiating L with respect to the variables $\boldsymbol{\omega}_1, b_1, \boldsymbol{\xi}, \varepsilon_1$, we have

$$\begin{aligned} \frac{\partial L}{\partial \boldsymbol{\omega}_1} &= -K\left(\mathbf{A}, \mathbf{A}^T\right)^T\mathbf{D}_1\left(\mathbf{y} - \left(K\left(\mathbf{A}, \mathbf{A}^T\right)\boldsymbol{\omega}_1 + b_1\mathbf{e}\right)\right) \\ &\quad + K\left(\mathbf{A}, \mathbf{A}^T\right)^T\boldsymbol{\alpha} + c_1\boldsymbol{\omega}_1 = 0, \end{aligned} \tag{15}$$

$$\begin{aligned} \frac{\partial L}{\partial b_1} &= -\mathbf{e}^T\mathbf{D}_1\left(\mathbf{y} - \left(K\left(\mathbf{A}, \mathbf{A}^T\right)\boldsymbol{\omega}_1 + b_1\mathbf{e}\right)\right) + \mathbf{e}^T\boldsymbol{\alpha} \\ &\quad + c_1 b_1 = 0, \end{aligned} \tag{16}$$

$$\frac{\partial L}{\partial \boldsymbol{\xi}} = c_2\mathbf{D}_2 - \boldsymbol{\alpha} - \boldsymbol{\beta} = 0, \tag{17}$$

$$\frac{\partial L}{\partial \varepsilon_1} = c_2\nu_1 - \mathbf{e}^T\boldsymbol{\alpha} - \gamma = 0, \tag{18}$$

The K.K.T. conditions are given by

$$\begin{aligned} \mathbf{y} - \left(K\left(\mathbf{A}, \mathbf{A}^T\right)\boldsymbol{\omega}_1 + b_1\mathbf{e}\right) &\geq -\varepsilon_1\mathbf{e} - \boldsymbol{\xi}, \\ \boldsymbol{\xi} &\geq 0\mathbf{e}, \\ \boldsymbol{\alpha}^T\left(\mathbf{y} - \left(K\left(\mathbf{A}, \mathbf{A}^T\right)\boldsymbol{\omega}_1 + b_1\mathbf{e}\right) + \varepsilon_1\mathbf{e} + \boldsymbol{\xi}\right) &= 0, \\ \boldsymbol{\alpha} &\geq 0\mathbf{e}, \\ \boldsymbol{\beta}^T\boldsymbol{\xi} &= 0, \\ \boldsymbol{\beta} &\geq 0\mathbf{e}, \\ \gamma\varepsilon_1 = 0, \quad \gamma &\geq 0. \end{aligned} \tag{19}$$

Combining (15) and (16), we obtain

$$\begin{aligned} &-\begin{bmatrix} K\left(\mathbf{A}, \mathbf{A}^T\right)^T \\ \mathbf{e}^T \end{bmatrix}\mathbf{D}_1\left(\mathbf{y} - \left(K\left(\mathbf{A}, \mathbf{A}^T\right)\boldsymbol{\omega}_1 + b_1\mathbf{e}\right)\right) \\ &+ \begin{bmatrix} K\left(\mathbf{A}, \mathbf{A}^T\right)^T \\ \mathbf{e}^T \end{bmatrix}\boldsymbol{\alpha} + c_1\begin{bmatrix} \boldsymbol{\omega}_1 \\ b_1 \end{bmatrix} = 0. \end{aligned} \tag{20}$$

Let $\mathbf{H} = [K(\mathbf{A}, \mathbf{A}^T)\mathbf{e}]$ and $\mathbf{u}_1 = [\boldsymbol{\omega}_1^T b_1]^T$, and (20) can be rewritten as

$$-\mathbf{H}^T\mathbf{D}_1\mathbf{y} + \left(\mathbf{H}^T\mathbf{D}_1\mathbf{H} + c_1\mathbf{I}\right)\mathbf{u}_1 + \mathbf{H}^T\boldsymbol{\alpha} = 0. \tag{21}$$

where $\mathbf{I}$ is an identity matrix with appropriate dimension. From (21), we get

$$\mathbf{u}_1 = \left(\mathbf{H}^T\mathbf{D}_1\mathbf{H} + c_1\mathbf{I}\right)^{-1}\mathbf{H}^T\left(\mathbf{D}_1\mathbf{y} - \boldsymbol{\alpha}\right). \tag{22}$$

From (15), (16), and (19), the following constraints can be obtained:

$$\mathbf{e}^T\boldsymbol{\alpha} \leq c_2\nu_1, \quad 0\mathbf{e} \leq \boldsymbol{\alpha} \leq c_2\mathbf{D}_2. \tag{23}$$

Substituting (22) into Lagrangian function L, we can get the dual formulation of (12)

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \quad &\frac{1}{2}\boldsymbol{\alpha}^T\mathbf{H}\left(\mathbf{H}^T\mathbf{D}_1\mathbf{H} + c_1\mathbf{I}\right)^{-1}\mathbf{H}^T\boldsymbol{\alpha} + \mathbf{y}^T\boldsymbol{\alpha} \\ &- \mathbf{y}^T\mathbf{D}_1\mathbf{H}\left(\mathbf{H}^T\mathbf{D}_1\mathbf{H} + c_1\mathbf{I}\right)^{-1}\mathbf{H}^T\boldsymbol{\alpha} \\ \text{s.t.:} \quad &0\mathbf{e} \leq \boldsymbol{\alpha} \leq c_2\mathbf{D}_2, \quad \mathbf{e}^T\boldsymbol{\alpha} \leq c_2\nu_1. \end{aligned} \tag{24}$$

Similarly, the dual formulation of (13) can be derived as follows:

$$\begin{aligned} \min_{\boldsymbol{\eta}} \quad &\frac{1}{2}\boldsymbol{\eta}^T\mathbf{H}\left(\mathbf{H}^T\mathbf{D}_1\mathbf{H} + c_3\mathbf{I}\right)^{-1}\mathbf{H}^T\boldsymbol{\eta} \\ &+ \mathbf{y}^T\mathbf{D}_1\mathbf{H}\left(\mathbf{H}^T\mathbf{D}_1\mathbf{H} + c_3\mathbf{I}\right)^{-1}\mathbf{H}^T\boldsymbol{\eta} - \mathbf{y}^T\boldsymbol{\eta} \\ \textit{st:} \quad &0\mathbf{e} \leq \boldsymbol{\eta} \leq c_4\mathbf{D}_2, \quad \mathbf{e}^T\boldsymbol{\eta} \leq c_4\nu_2. \end{aligned} \tag{25}$$

To solve the above QPPs, the vectors $\boldsymbol{\omega}_1, b_1$ and $\boldsymbol{\omega}_2, b_2$ can be expressed by

$$\begin{bmatrix} \boldsymbol{\omega}_1 \\ b_1 \end{bmatrix} = \left(\mathbf{H}^T\mathbf{D}_1\mathbf{H} + c_1\mathbf{I}\right)^{-1}\mathbf{H}^T\left(\mathbf{D}_1\mathbf{y} - \boldsymbol{\alpha}\right), \tag{26}$$

and

$$\begin{bmatrix} \boldsymbol{\omega}_2 \\ b_2 \end{bmatrix} = \left(\mathbf{H}^T\mathbf{D}_1\mathbf{H} + c_3\mathbf{I}\right)^{-1}\mathbf{H}^T\left(\mathbf{D}_1\mathbf{y} + \boldsymbol{\eta}\right). \tag{27}$$

By substituting the above results into (3), the final regression function can be obtained.

2.3.2. Determination of Wavelets Transform Based Weighted Matrix $\mathbf{D}_1$ and a Coefficient Vector $\mathbf{D}_2$. The wavelet transform can be used to denoise the time series signal. Based on the work of Ingrid Daubechies, the Daubechies wavelets

are a family of orthogonal wavelets for a discrete wavelet transform. There is a scaling function (called father wavelet) for each wavelet type, which generates an orthogonal multiresolution analysis. Daubechies orthogonal wavelets $db1$ – $db10$ are commonly used.

The wavelet transform process includes three stages: decomposition, signal processing, and reconstruction. In each step of decomposition, the signal can be decomposed into two sets of signals: high frequency signal and low frequency signal. Suppose that there is a time series signal $\mathbf{S}$. After the first step of decomposition, it generates two signals: one is high frequency part $\mathbf{Sh}_1$ and another is low frequency part $\mathbf{Sl}_1$. Then, in the second step of decomposition, the low frequency signal $\mathbf{Sl}_1$ can be decomposed further into $\mathbf{Sh}_2$ and $\mathbf{Sl}_2$. After k steps of decomposition, $k+1$ groups of decomposed sequence $(\mathbf{Sh}_1, \mathbf{Sh}_2, \ldots, \mathbf{Sh}_k, \mathbf{Sl}_k)$ are obtained, where $\mathbf{Sl}_k$ represents the contour of the original signal $\mathbf{S}$ and $\mathbf{Sh}_1, \mathbf{Sh}_2, \ldots, \mathbf{Sh}_k$ represent the subtle fluctuations. The decomposition process of signal $\mathbf{S}$ is shown in Figure 1 and defined as follows:

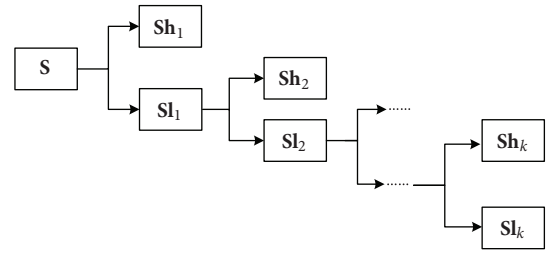


Figure 1: The process of decomposition.

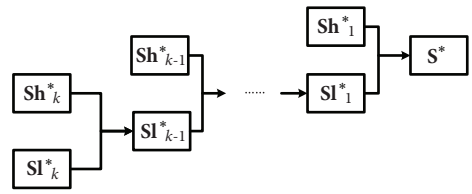


Figure 2: The process of reconstruction.

$$\mathbf{Sl}_k(n) = \sum_{m=1}^{l_{k-1}} \phi(m-2n)\,\mathbf{Sl}_{k-1}(m) \tag{28}$$

$$\mathbf{Sh}_k(n) = \sum_{m=1}^{l_{k-1}} \varphi(m-2n)\,\mathbf{Sl}_{k-1}(m) \tag{29}$$

where $\mathbf{Sl}_{k-1}$ with the length l_{k-1} is the signal to be decomposed, $\mathbf{Sl}_k$ and $\mathbf{Sh}_k$ are the results in the k-th step. $\phi(m-2n)$, and $\varphi(m-2n)$ are called scaling sequence (low pass filter) and wavelets sequence, respectively [24]. In this paper, $db2$ wavelet with the length of 4 is adopted. After wavelet transforming, appropriate signal processing can be carried out. In the stage of reconstruction, the processed high frequency signal $\mathbf{Sh}_k^*$ and low frequency signal $\mathbf{Sl}_k^*$ are reconstructed to generate the target signal $\mathbf{S}^*$. Let $\mathbf{P}$ and $\mathbf{Q}$ be the matrix with the order of $l_k \times l_{k-1}$, $\mathbf{P}_{n,m} = \phi(m-2n)$, and $\mathbf{Q}_{n,m} = \varphi(m-2n)$. The reconstruction process is shown in Figure 2. Therefore, (28) and (29) can be rewritten as follows:

$$\mathbf{Sl}_k(n) = \mathbf{P} \cdot \mathbf{Sl}_{k-1}(m) \tag{30}$$

$$\mathbf{Sh}_k(n) = \mathbf{Q} \cdot \mathbf{Sl}_{k-1}(m) \tag{31}$$

The signal $\mathbf{Sl}_{k-1}^*$ can be generated by reconstructing $\mathbf{Sl}_k^*$ and $\mathbf{Sh}_k^*$:

$$\mathbf{Sl}_{k-1}^*(m) = \mathbf{P}^{-1} \cdot \mathbf{Sl}_k^*(n) + \mathbf{Q}^{-1} \cdot \mathbf{Sh}_k^*(n) \tag{32}$$

If the output vector $\mathbf{S} = [S_1, S_2, \ldots, S_l]^T$ of a prediction model is a time sequence, then the wavelets transform can be used to denoise the output of the training samples. After the decomposition, signal processing, and reconstruction process, a denoising sequence $\mathbf{S}^* = [S_1^*, S_2^*, \ldots, S_l^*]^T$ can be obtained. The absolute difference vector $r_i = |S_i - S_i^*|$, $i = 1, 2, \ldots, l$, and $\mathbf{r} = [r_1, r_2, \ldots, r_l]^T$ between S and S^* denote the distance from the training samples to the denoising samples. In the WTWTSVR, a small, r_i, reflects a large weight on the distance between the estimated value and training value of the i-th sample, which can be defined as the following Gaussian function:

$$d_i = \exp\left(-\frac{r_i^2}{2\sigma^{*2}}\right), \quad i = 1, 2, \ldots, l, \tag{33}$$

where d_i denotes the weight coefficient and σ^* is the width of the Gaussian function. Therefore, the wavelets transform based weighted matrix $\mathbf{D}_1$ and the coefficient vector $\mathbf{D}_2$ can be determined by $\mathbf{D}_1 = \mathrm{diag}(d_1, d_2, \ldots, d_l)$ and $\mathbf{D}_2 = [d_1, d_2, \ldots, d_l]^T$.

2.3.3. Computational Complexity Analysis. The computation complexity of the proposed algorithm is mainly determined by the computations of a pair of QPPs and a pair of inverse matrices. If the number of the training samples is l, then the training complexity of dual QPPs is about $O(2l^3)$, while the training complexity of the traditional SVR is about $O(8l^3)$, which implies that the training speed of SVR is about four times the proposed algorithm. Also, a pair of inverse matrices with the size $(l+1) \times (l+1)$ in QPPs have the same computational cost $O(l^3)$. During the training process, it is a good way to cache the pair of inverse matrices with some memory cost in order to avoid repeated computations. In addition, the proposed algorithm contains the wavelet transform weighted matrix and $db2$ wavelet with length of 4 is used in this paper. Then, the complexity of wavelet transform is less than $8l$. By comparing with the computations of QPPs and inverse matrix, the complexity of computing the wavelet matrix can be ignored. Therefore, the computation complexity of the proposed algorithm is about $O(3l^3)$.

3. Establishment of Static Control Model

The end-point carbon content (denoted as C) and the temperature (denoted as T) are main factors to test the quality of steelmaking. The ultimate goal of steelmaking is to control

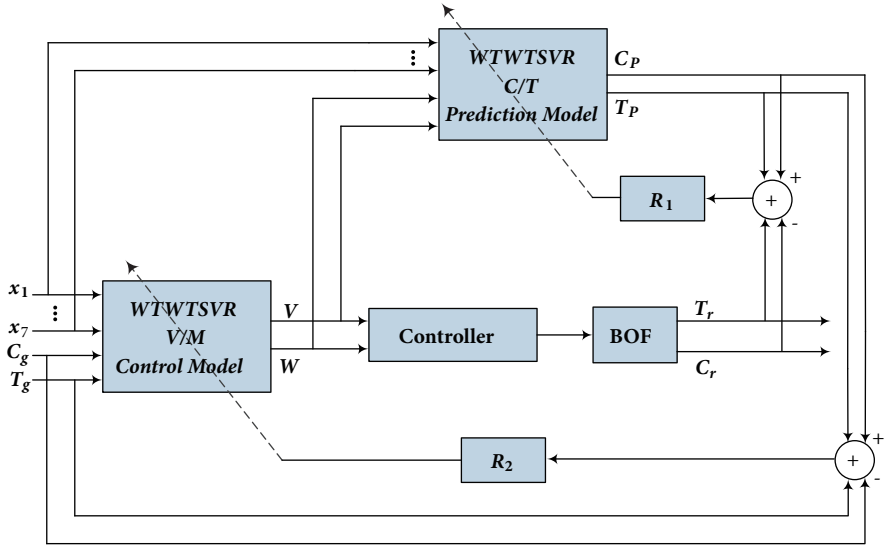


Figure 3: Structure of proposed BOF control model.

C and T to a satisfied region. BOF steelmaking is a complex physicochemical process, so its mathematical model is very difficult to establish. Therefore, the intelligent method such as TSVR can be used to approximate the BOF model. From the collected samples, it is easy to see that the samples are listed as a time sequence, which means they can be seen as a time sequence signal. Hence, the proposed WTWTSVR algorithm is appropriate to establish a model for BOF steelmaking.

In this section, a novel static control model for BOF is established based on WTWTSVR algorithm. According to the initial conditions of the hot metal and the desired end-point carbon content (denoted as C_g) and the temperature (denoted as T_g), the relative oxygen blowing volume (denoted as V) and the weight of auxiliary raw material (denoted as W) can be calculated by the proposed model. Figure 3 shows the structure of the proposed BOF control model, which is composed of a prediction model for carbon content and temperature (C/T prediction model), a control model for oxygen blowing volume and the weight of auxiliary raw materials (V/W control model), two parameter regulating units (R_1 and R_2), and a controller and a basic oxygen furnace in any plant. Firstly, the WTWTSVR C/T prediction model should be established by using the historic BOF samples, which consists of two individual models (C_model and T_model). C_model represents the prediction model of the end-point carbon content and T_model represents the prediction of the end-point temperature. The inputs of two models both include the initial conditions of the hot metal, and the outputs of them are C and T, respectively. The parameters of the prediction models can be regulated by R_1 to obtain the optimized models. Secondly, the WTWTSVR V/W control model should be established based on the proposed prediction models, and the oxygen blowing volume can be calculated by V_model and the weight of auxiliary raw materials can be determined by W_model. The inputs of the control models both include the initial conditions of the hot metal, the desired end-point carbon content C_g), and the end-point temperature T_g. The outputs of them are V and W, respectively. The parameters of the control models can be regulated by R_2. After the regulation of R_1 and R_2, the static control model can be established. For any future hot metal, the static control model can be used to calculate V and W by the collected initial conditions and C_g and T_g. Then, the calculated values of V and W will be sent to the controller. Finally, the relative BOF system is controlled by the controller to reach the satisfactory end-point region.

3.1. Establishment of WTWTSVR C/T Prediction Model. To realize the static BOF control, an accurate prediction model of BOF should be designed firstly. The end-point prediction model is the foundation of the control model. By collecting the previous BOF samples, the abnormal samples must be discarded, which may contain the wrong information of BOF. Through the mechanism analysis of BOF, the influence factors on the end-point information are mainly determined by the initial conditions of hot metal, which means the influence factors are taken as the independent input variables of the prediction models. The relative input variables are listed in Table 1. Note that the input variable x_9 denotes the sum of the weight of all types of auxiliary materials, which are including the light burned dolomite, the dolomite stone, the lime ore and scrap, etc. It has a guiding significance for real production, and the proportion of the components can be determined by the experience of the user. The output variable of the model is the end-point carbon content C or the end-point temperature T.

According to the prepared samples of BOF and WTWTSVR algorithm, the regression function $f_{C/T}(x)$ can be given by

$$f_{C/T}(\mathbf{x}) = \frac{1}{2}K\left(\mathbf{x}^T, \mathbf{A}^T\right)\left(\boldsymbol{\omega}_1 + \boldsymbol{\omega}_2\right)^T + \frac{1}{2}\left(b_1 + b_2\right) \quad (34)$$

where $f_{C/T}(\mathbf{x})$ denotes the estimation function of C_model or T_model and $\mathbf{x} = [x_1, x_2, \ldots, x_9]^T$.

Table 1: Independent input variables of prediction models.

Name of input variable	Symbol	Units	Name of input variable	Symbol	Units
Initial carbon content	x_1	%	Sulphur content	x_6	%
Initial temperature	x_2	˚C	Phosphorus content	x_7	%
Charged hot metal	x_3	*tons*	Oxygen blowing volume	$x_8\ (V)$	Nm^3
Silicon content	x_4	%	Total weight of auxiliary raw materials	$x_9\ (W)$	*tons*
Manganese content	x_5	%			

Out of the historical data collected from one BOF of 260 *tons* in some steel plant in China, 220 samples have been selected and prepared. In order to establish the WTWTSVR prediction models, the first 170 samples are taken as the training data and 50 other samples are taken as the test data to verify the accuracy of the proposed models. In the regulating unit R_1, the appropriate model parameters ($c_1, c_2, c_3, c_4, \nu_1, \nu_2$, and σ_1, σ_1^*) are regulated manually, and then the C_model and T_model can be established. In summary, the process of the modelling can be described as follows.

Step 1. Initialize the parameters of the WTWTSVR prediction model, and normalize the prepared 170 training samples from its original range to the range [-1 1] by *mapminmax* function in Matlab.

Step 2. Denoise the end-point carbon content $\mathbf{C} = [C_1, C_2, \ldots, C_l]^T$ or temperature $\mathbf{T} = [T_1, T_2, \ldots, T_l]^T$ in the training samples by using the wavelet transform described in the previous section. Then, the denoised samples $\mathbf{C}^* = [C_1^*, C_2^*, \ldots, C_l^*]^T$ and $\mathbf{T}^* = [T_1^*, T_2^*, \ldots, T_l^*]^T$ can be obtained.

Step 3. By selecting the appropriate parameter σ_1^* in (33), determine the wavelets transform based weighted matrix $\mathbf{D}_1$ and the coefficient vector $\mathbf{D}_2$.

Step 4. Select appropriate values of $c_1, c_2, c_3, c_4, \nu_1, \nu_2$ and σ_1 in the regulating unit R_1.

Step 5. Solve the optimization problems in (24) and (25) by *quadprog* function in Matlab and return the optimal vector $\boldsymbol{\alpha}$ or $\boldsymbol{\eta}$.

Step 6. Calculate the vectors $\boldsymbol{\omega}_1, b_1$ and $\boldsymbol{\omega}_2, b_2$ by (26) and (27).

Step 7. Substitute the parameters in Step 5 into (3) to obtain the function $f_{C/T}(\mathbf{x})$.

Step 8. Substitute the training samples into (34) to calculate the relative criteria of the model, which will be described in details later.

Step 9. If the relative criteria are satisfactory, then the C_model or T_model is established. Otherwise, return to Steps from 4 to 8.

3.2. Establishment of WTWTSVR V/W Control Model. Once the WTWTSVR C/T prediction models are established, the WTWTSVR V/W control models are ready to build. In order to control the end-point carbon content and temperature to the satisfactory region for any future hot metal, the relative oxygen blowing volume V and the total weight of auxiliary raw materials W should be calculated based on V/W control models, respectively. Through the mechanism analysis, the influence factors on V and W are mainly determined by the initial conditions of hot metal, and the desired conditions of the steel are also considered as the the influence factors. Then, the independent input variables of the control models are listed in Table 2, where the input variables x_1-x_7 are the same as the C/T prediction model and other two input variables are the desired carbon content C_g and desired temperature T_g, respectively. The output variable of the control model is V or W defined above.

Similar to (34), the regression function $f_{V/W}(\mathbf{x})$ can be written as

$$f_{V/W}(\mathbf{x}) = \frac{1}{2}K\left(\mathbf{x}^T, \mathbf{A}^T\right)\left(\boldsymbol{\omega}_3 + \boldsymbol{\omega}_4\right)^T + \frac{1}{2}\left(b_3 + b_4\right) \quad (35)$$

where $f_{V/W}(\mathbf{x})$ denotes the estimation function of V_model or W_model and $\mathbf{x} = [x_1, x_2, \ldots, x_9]^T$.

In order to establish the WTWTSVR control models, the training samples and the test samples are the same as that of the prediction models. The regulating unit R_2 is used to regulate the appropriate model parameters ($c_5, c_6, c_7, c_8, \nu_3, \nu_4$ and σ_2, σ_2^*) manually; then the V_model and W_model can be established. In summary, the process of the modelling can be described as follows.

Steps 1–7. Similar to those in the section of establishing the prediction models except for the input variables listed in Table 2 and the output variable being the oxygen blowing volume $\mathbf{V} = [V_1, V_2, \ldots, V_l]^T$ or the total weight of auxiliary materials $\mathbf{W} = [W_1, W_2, \ldots, W_l]^T$. Also, select appropriate values of $c_5, c_6, c_7, c_8, \nu_3, \nu_4$ and σ_2 in the regulating unit R_2; then the function $f_{V/W}(\mathbf{x})$ is obtained.

Step 8. Substitute the training samples into (35) to calculate the relative predicted values $\widehat{V}$ and $\widehat{W}$ of V and W respectively.

Step 9. $\widehat{V}$ and $\widehat{W}$ are combined with the input variables $x_1 - x_7$ in Table 1 to obtain 50 new input samples. Then, substituting them into the C_model and T_model, 50 predicted values $\widehat{C}$ and $\widehat{T}$ can be obtained. Finally, calculate the differences between the desired values C_g, T_g and the predicted values $\widehat{C}, \widehat{T}$. Also, other relative criteria of the model should be determined.

Table 2: Independent input variables of control models.

Name of input variable	Symbol	Units	Name of input variable	Symbol	Units
Initial carbon content	x_1	%	Sulphur content	x_6	%
Initial temperature	x_2	°C	Phosphorus content	x_7	%
Charged hot metal	x_3	*tons*	Desired carbon content	x_8 (C_g)	Nm^3
Silicon content	x_4	%	Desired temperature	x_9 (T_g)	*tons*
Manganese content	x_5	%			

Step 10. If the obtained criteria are satisfactory, then the V_model or W_model is established. Otherwise, return to Steps from 4 to 9.

4. Results and Discussion

In order to verify the performances of WTWTSVR algorithm and proposed models, the artificial functions and practical datasets are adopted, respectively. All experiments are carried out in Matlab R2011b on Windows 7 running on a PC with Intel (R) Core (TM) i7-4510U CPU 2.60GHz with 8GB of RAM.

The evaluation criteria are specified to evaluate the performance of the proposed method. Assume that the total number of testing samples is n, y_i is the actual value at the sample point x_i, $\widehat{y}_i$ is the estimated value of y_i, and $\overline{y}_i = (\sum_{i=1}^{n} y_i)/n$ is the mean value of $y_1, y_2, \ldots, y_n$. Therefore, the following criteria can be defined:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}, \tag{36}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \widehat{y}_i\right|, \tag{37}$$

$$\frac{SSE}{SST} = \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y}_i)^2}, \tag{38}$$

$$\frac{SSR}{SST} = \frac{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y}_i)^2}, \tag{39}$$

$$HR = \frac{Number\ of\ \left|y_i - \widehat{y}_i\right| < error\ bound}{n} \times 100\%, \tag{40}$$

In the above equations, *RMSE* denotes the root mean squared error, *MAE* denotes the mean absolute error, and *SSE*, *SST*, and *SSR* represent the sum of the squared deviation between any two of $y_i, \widehat{y}_i, \overline{y}_i$, respectively. Normally, a smaller value of *SSE*/*SST* reflects that the model has a better performance. The decrease in *SSE*/*SST* reflects the increase in *SSR*/*SST*. However, if the value of *SSE*/*SST* is extremely small, it will cause the overfitting problem of the regressor. An important criteria for evaluating the performance of the BOF model is hit rate (*HR*) calculated by (40), which is defined as the ratio of the number of the satisfactory samples over the total number of samples. For any sample in the dataset, if the absolute error between the estimated value and actual value is smaller than a certain error bound, then the results of the sample hit the end point. A large number of hit rate indicate a better performance of the BOF model. Generally, a hit rate of 90% for C or T is a satisfactory value in the real steel plants.

4.1. Performance Test of WTWTSVR. In this section, the artificial function named Sinc function is used to test the regression performance of the proposed WTWTSVR method, which can be defined as $y = \sin x/x,\ x \in [-4\pi, 4\pi]$.

In order to evaluate the proposed method effectively, the training samples are polluted by four types of noises, which include the Gaussian noises with zero means and the uniformly distributed noises. For the train samples $(x_i, y_i),\ i = 1, 2, \ldots, l$, we have

$$y_i = \frac{\sin x_i}{x_i} + \theta_i, \quad x \sim U[-4\pi, 4\pi],\ \theta_i \sim N\left(0, 0.1^2\right), \tag{41}$$

$$y_i = \frac{\sin x_i}{x_i} + \theta_i, \quad x \sim U[-4\pi, 4\pi],\ \theta_i \sim N\left(0, 0.2^2\right), \tag{42}$$

$$y_i = \frac{\sin x_i}{x_i} + \theta_i, \quad x \sim U[-4\pi, 4\pi],\ \theta_i \sim U[0, 0.1], \tag{43}$$

$$y_i = \frac{\sin x_i}{x_i} + \theta_i, \quad x \sim U[-4\pi, 4\pi],\ \theta_i \sim U[0, 0.2], \tag{44}$$

where $U[m, n]$ and $N(p, q^2)$ denote the uniformly random variable in $[m, n]$ and the Gaussian variable with the mean p and variance q^2, respectively. For all regressions of the above four Sinc functions, 252 training samples and 500 test samples are selected. Note that the test samples are uniformly sampled from the Sinc functions without any noise. For all algorithms in this paper, the following sets of parameters are explored: the penalty coefficient c is searched from the set $\{i/1000, i/100, i/10 \mid i = 1, 2, \ldots, 10\}$. The tube regulation coefficient v is selected from the set $\{1, 2, \ldots, 10\}$. The wavelet coefficient σ and kernel parameter σ^* are both searched over the range $\{i/1000, i/100, i/10, i \mid i = 1, 2, \ldots, 10\}$. In order to reduce the number of combinations in the parameter search, the following relations are chosen: $c_1 = c_3$, $c_2 = c_4$, and $v_1 = v_2$.

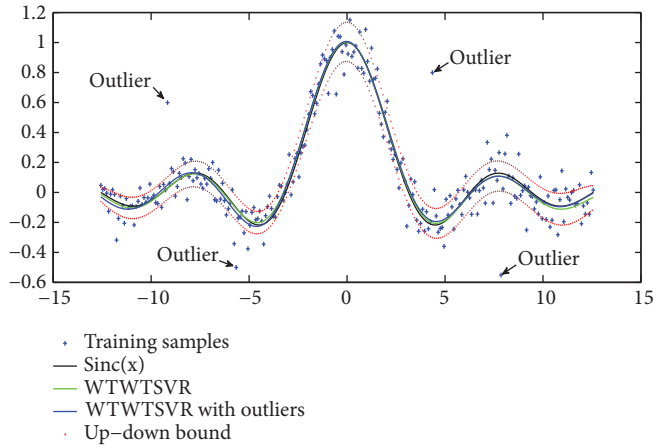


Figure 4: Performance of WTWTSVR on Sinc function with and without outliers.

Firstly, the choice of kernel function should be considered, RBF kernel is an effective and frequently used kernel function in TSVR research papers. Also, the performance of RBF has been evaluated by comparing with other three existing kernel functions (listed in Table 3), and the results are shown in Table 4. It is easy to see that the RBF kernel achieves the optimal results. Then, the average results of the proposed methods and other four existing methods (TSVR [15], ν-TSVR [17], KNNWTSVR [19] and Asy ν-TSVR [20]) with 10 independent runs are shown in Table 5, where Type A, B, C and D denote four different types of noises (41)-(44). Obviously, the results of Table 5 show that the proposed method achieves the optimal result of *SSE*. Also, the proposed method achieves the smallest *SSE*/*SST* results in four regressions of Sinc functions, which are 0.0029, 0.0124, 0.0231 and 0.0903. Especially, the obvious performances are enhanced in Type A and B. The *SSR*/*SST* of the proposed method takes the first position in Type B, the second position in Type A and C, and the third position in Type D. From the aspect of training time, it can be seen that the proposed method achieves the optimal result in Type B. In Type A, C and D, the training time of the proposed method is faster than that of TSVR and KNNWTSVR, and close to ν-Type and Asy ν-Type. It verifies that the computational complexity of wavelet transform weighting scheme is lower than KNN weighting scheme. For the effect of outliers, we have verified the performance of the proposed method on the Sinc function with Type A noise. Figure 4 shows that the prediction performance of the proposed method is satisfied against the outliers, as shown in the blue line, and the results of *SSE*, *SSE*/*SST* and *SSR*/*SST* achieve 0.1911, 0.0036 and 1.0372, respectively. By comparing with the results in Table 5, it can be concluded that the overall performance of the proposed method with outliers is still better than TSVR, v-TSVR and KNNWTSVR. Therefore, it can be concluded that the overall regression performance of the proposed method is optimal.

4.2. Verification of the Proposed BOF Models. In order to verify the effectiveness of the proposed model, 240 heats samples for low carbon steel of 260*tons* BOF were collected from some steel plant in China. Firstly, the preprocessing of the samples should be carried out. The abnormal samples with the unexpected information must be removed, which exceeds the actual range of the signals. It may be caused by the wrong sampling operation. Then, 220 qualified samples are obtained. The next step is to analyze the information of the qualified samples. The end-point carbon content and temperature are mainly determined by the initial conditions of the iron melt, the total blowing oxygen volume and the weight of material additions. Therefore, the unrelated information should be deleted, such as the heat number, the date of the steelmaking and so on. According to the information of Tables 1 and 2, the datasets for the relative prediction and control models can be well prepared. After that, the proposed models are ready to be established. More details of the model have been described in Section 3. The proposed control model consists of two prediction models (C_model and T_model) and two control models (V_model and W_model). For each individual model, there are 8 parameters that need to be regulated to achieve the satisfactory results. The regulation processes are related to the units R_1 and R_2 in Figure 3. First of all, the prediction models need to be established. In order to meet the requirements of the real production, the prediction errors bound with 0.005% for C_model and 10°C for T_model are selected. Similarly, the control errors bound with 800 Nm^3 for V_model and 5.5 *tons* for W_model are selected. The accuracy of each model can be reflected by the hit rate with its relative error bound, and a hit rate of 90% within each individual error bound is a satisfied result.

Table 3: Type and functions of four existing kernel functions.

Kernel Type	Function
Radial Basis Function Kernel (RBF)	$K(\mathbf{x}, \mathbf{x}_i) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}_i\|^2}{2\sigma^2}\right)$
Rational Quadratic Kernel (RQ)	$K(\mathbf{x}, \mathbf{x}_i) = 1 - \frac{\|\mathbf{x}-\mathbf{x}_i\|^2}{\left(\|\mathbf{x}-\mathbf{x}_i\|^2 + \sigma\right)}$
Multiquadric Kernel	$K(\mathbf{x}, \mathbf{x}_i) = \left(\|\mathbf{x}-\mathbf{x}_i\|^2 + \sigma^2\right)^{0.5}$
Log Kernel	$K(\mathbf{x}, \mathbf{x}_i) = -\log\left(1 + \|\mathbf{x}-\mathbf{x}_i\|^{\sigma}\right)$

Table 4: . Comparisons of WTWTSVR on Sinc functions with different kernel functions.

Noise	Kernel Type	*SSE*	*SSE/SST*	*SSR/SST*
Type A	RBF	**0.1577±0.0099**	**0.0029±0.0018**	**0.9972±0.0264**
	RQ	0.1982±0.0621	0.0037±0.0012	1.0157±0.0297
	Multiquadric	0.2884±0.0762	0.0054±0.0014	0.9911±0.0345
	Log	0.4706±0.1346	0.0088±0.0025	1.0115±0.0306
Type B	RBF	**0.6659±0.2456**	**0.0124±0.0046**	**1.0161±0.0564**
	RQ	1.1662±0.3175	0.0217±0.0059	0.9834±0.0385
	Multiquadric	1.1101±0.5028	0.0206±0.0093	1.0030±0.0867
	Log	1.8157±0.6026	0.0338±0.0112	1.0366±0.0802
Type C	RBF	**1.2425±0.1444**	**0.0231±0.0027**	**1.0228±0.0153**
	RQ	3.5697±1.5023	0.0664±0.0279	1.0421±0.1584
	Multiquadric	3.6438±1.4810	0.0678±0.0275	1.0494±0.0641
	Log	6.9962±2.2691	0.1301±0.0422	1.1208±0.1682
Type D	RBF	**4.8557±0.4009**	**0.0903±0.0075**	**1.0768±0.0283**
	RQ	6.1541±1.8766	0.1144±0.0349	1.0829±0.1923
	Multiquadric	5.6597±2.4049	0.1052±0.0447	1.1015±0.1908
	Log	14.9196±3.4583	0.2774±0.0643	1.3255±0.2159

Table 5: Comparisons of four regression methods on Sinc functions with different noises.

Noise	Regressor	*SSE*	*SSE/SST*	*SSR/SST*	Time, *s*
Type A	WTWTSVR	**0.1577±0.0099**	**0.0029±0.0018**	0.9972±0.0264	1.7997
	TSVR	0.2316±0.1288	0.0043±0.0024	1.0050±0.0308	2.3815
	v-TSVR	0.4143±0.1261	0.0077±0.0023	0.9501±0.0270	1.7397
	Asy *v*-TSVR	0.1742±0.1001	0.0032±0.0019	1.0021±0.0279	1.7866
	KNNWTSVR	0.2044±0.0383	0.0038±0.0007	1.0177±0.0202	3.5104
Type B	WTWTSVR	**0.6659±0.2456**	**0.0124±0.0046**	**1.0161±0.0564**	**1.8053**
	TSVR	0.8652±0.3006	0.0161±0.0056	1.0185±0.0615	2.3296
	v-TSVR	0.8816±0.3937	0.0164±0.0073	0.9631±0.0548	1.8582
	Asy *v*-TSVR	0.7900±0.2588	0.0147±0.0048	1.0168±0.0599	1.8450
	KNNWTSVR	0.6891±0.2571	0.0128±0.0048	0.9679±0.0392	3.6175
Type C	WTWTSVR	**1.2425±0.1444**	**0.0231±0.0027**	1.0228±0.0153	1.8010
	TSVR	1.2505±0.0650	0.0233±0.0012	1.0217±0.0091	2.3054
	v-TSVR	1.5351±0.1172	0.0285±0.0022	0.9750±0.0111	1.7888
	Asy *v*-TSVR	1.2464±0.0934	0.0232±0.0017	1.0256±0.0125	1.6888
	KNNWTSVR	2.2027±0.7245	0.0410±0.0135	1.0148±0.1556	4.1230
Type D	WTWTSVR	**4.8557±0.4009**	**0.0903±0.0075**	1.0768±0.0283	1.7803
	TSVR	5.0090±0.2172	0.0931±0.0040	1.0890±0.0131	2.2556
	v-TSVR	5.2580±0.2935	0.0978±0.0055	1.0386±0.0125	1.7578
	Asy *v*-TSVR	4.9659±0.2925	0.0923±0.0054	1.0889±0.0128	1.7441
	KNNWTSVR	4.9751±1.3262	0.0925±0.0274	1.0262±0.1538	4.4397

Also, an end-point double hit rate (*DHR*) of the prediction model is another important criterion for the real production, which means the hit rates of end-point carbon content and temperature are both hit for the same sample. Hence, 170 samples are used to train the prediction and control models, and 50 samples are adopted to test the accuracy of the models. For each run of the relative modelling, the results are returned to the user with the information of the evaluation criteria. It shows the accuracy and fitness of the models with the specific parameters. Smaller values of *RMSE*, *MAE*, and *SSE*/*SST* and larger values of *SSR*/*SST* and hit rate are preferred, which means the model has a better generalization and higher accuracy. The regulation units R_1 and R_2 are used to balance the above criteria by selecting the appropriate values of the parameters. Note that the principle of the parameter selection is satisfactory if the following results are obtained: $SSR/SST > 0.5$ and $HR > 85\%$ with the smallest *SSE*/*SST*. Especially, the double hit rate *DHR* should be greater than 80 or higher. For

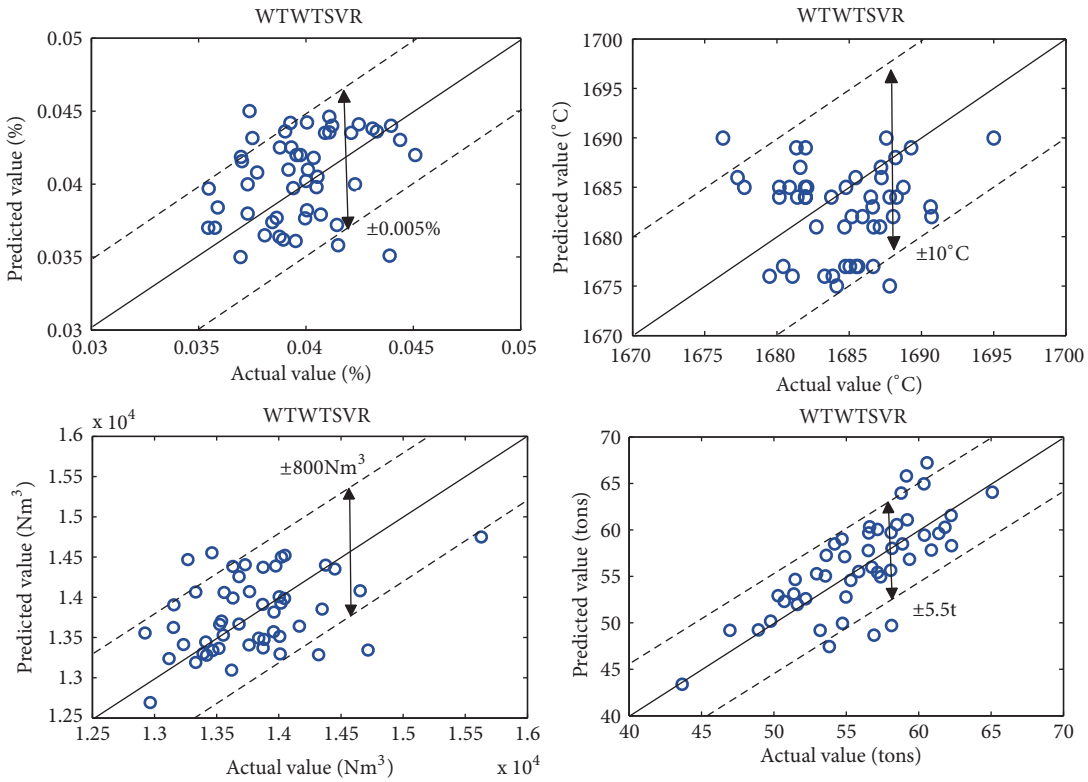


Figure 5: Performance of errors between predicted values and actual values with proposed static control model.

the collected samples, the parameters of WTWTSVR models are specified and shown in Table 6.

By using these parameters, the proposed BOF control model has been established. In order to evaluate the performance of the proposed method, more simulations of other existing regression methods with the samples are carried out, which are TSVR, v-TSVR, Asy v-TSVR, and KNNWTSVR, respectively. The comparison results of the carbon content prediction are listed in Table 7. From the results, it can be seen that the results of *RMSE*, *MAE*, and *SSE*/*SST* in the proposed method are 0.0023, 0.0026, and 1.1977, respectively. They are all smaller than those of other three existing methods, and the *SSR*/*SST* of 0.6930 is the highest result. The error performance and distribution of end-point carbon content between the predicted values and actual values are shown in Figures 5 and 6. It is clear that the proposed C_model can achieve the hit rate of 92%, which is better than other four methods. From above analysis, it illustrates that the proposed C_model has the best fitting behaviour for carbon content prediction.

Similarly, the performance comparisons of the temperature prediction are also listed in Table 7. From the results of Table 7, the best results of *RMSE* and *SSE*/*SST* and second best result of *MAE* of the proposed method are obtained. The results of *SSR*/*SST* is in the third position. Figures 5 and 6 show that the proposed T_model can achieve the hit rate of 96%, which is the optimal results by comparing with other methods. In addition, the double hit rate is also the key criterion in the real BOF applications; the proposed method can achieve a double hit rate of 90%, which is the best result, although the temperature hit rate of v-TSVR and KNNWTSVR method can also achieve 96%. However, their double hit rates only achieve 80% and 84%, respectively. In the real applications, a double hit rate of 90% is satisfactory. Therefore, the proposed model is more efficient to provide a reference for the real applications. Also, it meets the requirement of establishing the static control model.

Based on the prediction models, the control models (V_model and W_model) can be established by using the relative parameters in Table 6. The performance of the proposed V_model is shown in Figure 7. By comparing the proposed method with the existing methods, the comparison results of the oxygen blowing volume calculation are listed in Table 8, which shows that the results of *RMSE*, *MAE*, and *SSE*/*SST* in the proposed method are 371.3953, 411.7855, and 1.2713, respectively. They are all smaller than those of other four existing methods, and the *SSR*/*SST* of 1.0868 is in the fourth position. Figures 5 and 6 show that the predicted values of the proposed model agree well with the actual values of oxygen volume, and the proposed V_model has a best hit rate of 90%. It verifies that the proposed V_model has the best fitting behaviour for the calculation of oxygen blowing volume.

Similarly, the performance of the proposed W_model is shown in Figure 8, and the performance comparisons of the weight of the auxiliary materials are also listed in Table 8.

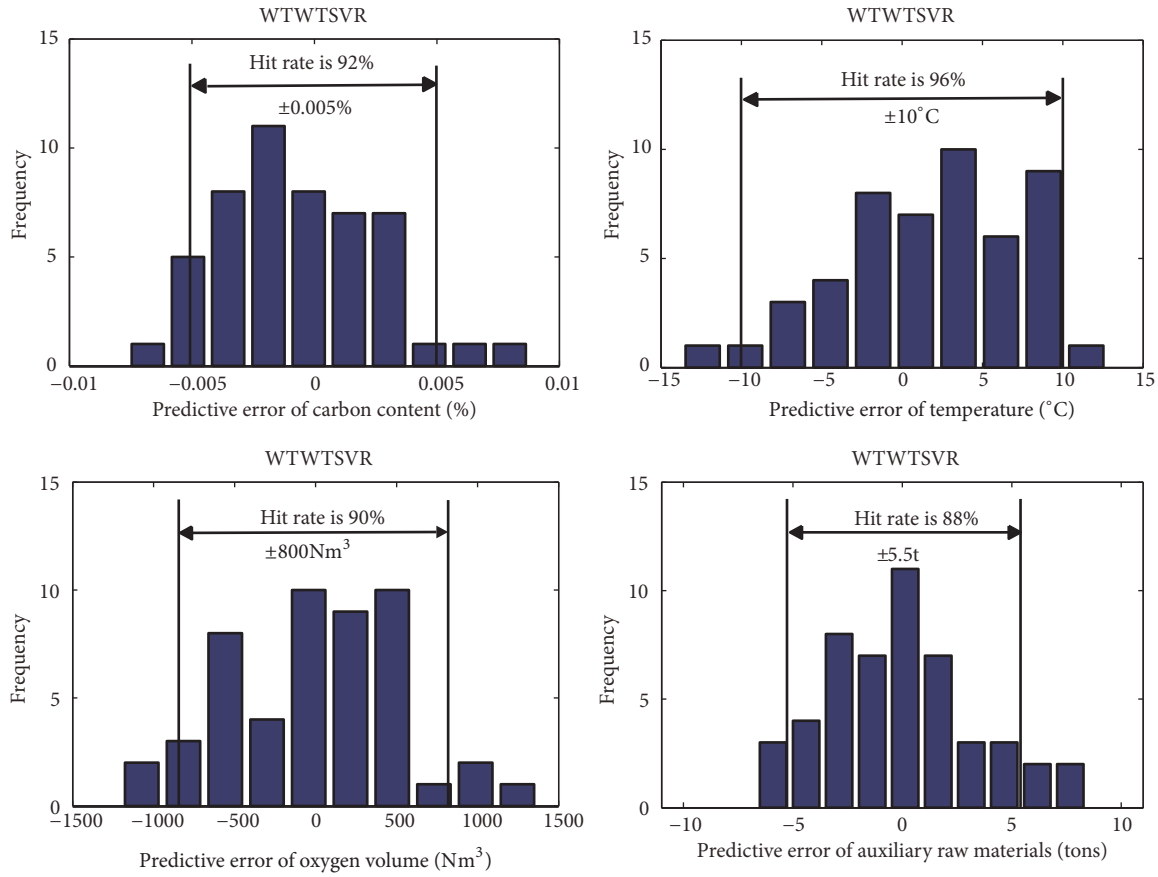


Figure 6: Performance of error distributions with proposed static control model.

Table 6: Specified parameters of prediction and control models.

Prediction Model	$c_1 = c_3$	$c_2 = c_4$	$v_1 = v_2$	σ_1	σ_1^*
C_model	0.005	0.02	3	0.005	1
T_model	0.002	0.05	5	0.04	0.3
Control Model	$c_5 = c_7$	$c_6 = c_8$	$v_3 = v_4$	σ_2	σ_2^*
V_model	0.002	0.01	2	0.2	3
W_model	0.001	0.02	1	0.5	5

Table 7: Performance comparisons of C/T prediction models with four methods.

Model	Criteria	WTWTSVR	TSVR	v-TSVR	Asy v-TSVR	KNNWTSVR
C_model (±0.005 %)	*RMSE*	**0.0023**	0.0026	0.0026	0.0026	0.0025
	MAE	**0.0026**	0.0031	0.0031	0.0031	0.0030
	SSE/*SST*	**1.1977**	1.6071	1.5675	1.5214	1.4680
	SSR/*SST*	**0.6930**	0.4616	0.3514	0.3835	0.3500
	HR,%	**92**	82	84	84	88
	Time, *s*	0.4523	0.2915	0.3706	0.3351	0.5352
T_model (±10 °C)	*RMSE*	**4.0210**	4.2070	4.3630	4.1013	4.0272
	MAE	4.7380	5.0363	5.1461	4.7970	4.7165
	SSE/*SST*	**1.7297**	1.8935	2.0365	1.7996	1.7351
	SSR/*SST*	0.7968	0.6653	0.7578	0.9141	0.8338
	HR,%	**96**	94	92	96	96
	Time, *s*	0.0977	0.1181	0.0861	0.0538	0.2393
	DHR,%	**90**	78	78	80	84

Table 8: Performance comparisons of V/W control models with four methods.

Model	Criteria	WTWTSVR	TSVR	v-TSVR	Asy v-TSVR	KNNWTSVR
V_model (±800 Nm^3)	*RMSE*	**371.3953**	383.0249	383.2387	399.8635	399.7568
	MAE	**411.7855**	416.3151	423.0260	427.0896	427.0415
	SSE/*SST*	**1.2713**	1.3522	1.3537	1.4737	1.4729
	SSR/*SST*	1.0868	1.1288	0.9691	0.9326	0.9319
	HR, %	**90**	86	84	86	86
	Time, *s*	0.6196	0.4397	0.5177	0.4883	0.7014
W_model (±5.5 *tons*)	*RMSE*	**2.4158**	2.8824	2.8431	3.6781	2.9077
	MAE	**2.7057**	3.1254	3.1632	3.9979	3.2588
	SSE/*SST*	**0.3791**	0.5398	0.5251	0.8789	0.5493
	SSR/*SST*	0.6505	0.5868	0.6739	0.4376	0.6725
	HR, %	**88**	86	80	80	84
	Time, *s*	0.1269	0.1230	0.0837	0.0609	0.2999
Hit rates	*HR* (*C*)	86	82	88	90	86
	HR (*T*)	92	94	94	80	94
	DHR	82	78	82	82	82

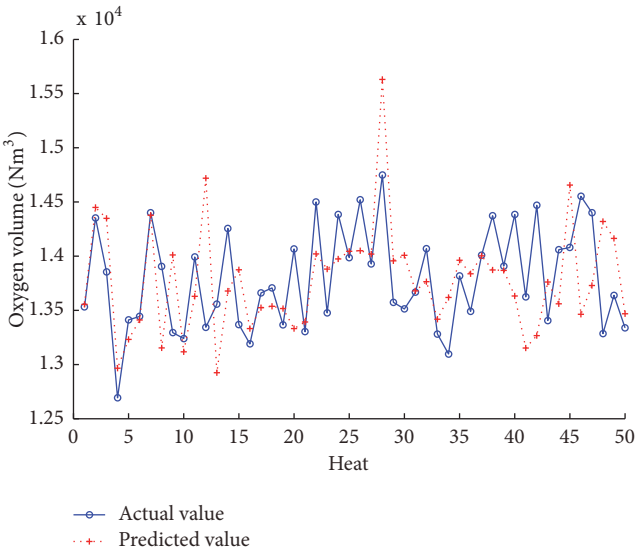


Figure 7: Performance of proposed V_model.

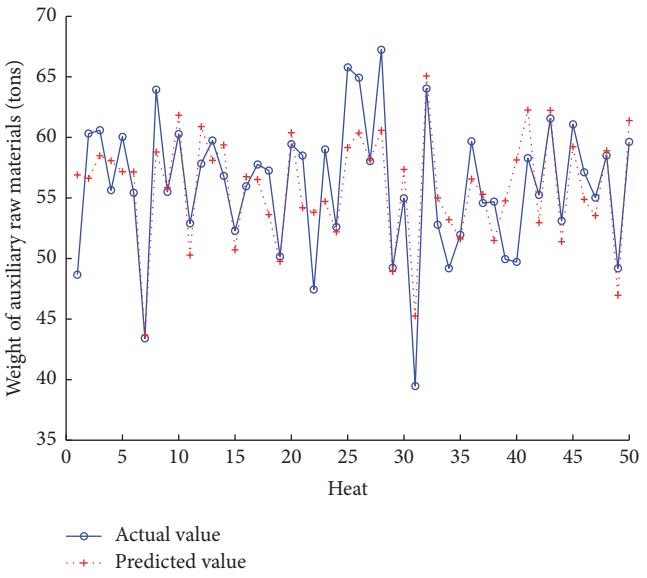


Figure 8: Performance of proposed W_model.

The proposed model achieves the best results of *RMSE*, *MAE*, and *SSE*/*SST*. Figures 5 and 6 show that the proposed method achieves the best hit rate of 88%. In addition, the training time of the proposed models is faster than that of KNNWTSVR method and slower than that of other three methods. It illustrates that the weighting scheme takes more time to obtain higher accuracy, and the performance of proposed weighting scheme is better than that of KNN weighting scheme for time sequence samples. For 50 test samples, there are 50 calculated values of V and W by V_model and W_model. Then, they are taken as the input variables into the proposed C_model and T_model to verify the end-point hit rate. From the results in Table 8, the proposed models can achieve a hit rate of 86% in C, 92% in T, and 82% in *DHR* is in the optimal result and the same as that of v-TSVR, Asy v-TSVR, and KNNWTSVR. In the real productions, *DHR* is paid more attention rather than the individual *HR* in C or T. Therefore, the proposed control models are verified to be able to guide the real production.

Based on above analysis, it can be concluded that the proposed static control model is effective and feasible; the hit rate can meet the requirements of the real productions for low carbon steel. For other types of steels, the proposed model is still suitable for use. Firstly, the specific samples of the heats should be obtained and preprocessed, and the analysis of the influence factors on the specific type of steel should be carried out to obtain the input variables of the model. The output variables are the same as the relative proposed models. Secondly, the end-point criteria should be specified, which is determined by the type of steel. Then, the parameters of the relative models can be regulated and determined to achieve the best criteria. Finally, the BOF control model for

the specific type of steel is established to guide the production in the plant.

BOF steelmaking is a complex physicochemical process; the proposed static control model can be established based on the real samples collected from the plant. However, there must be some undetected factors during the steelmaking process, which will affect the accuracy of the calculations of V and W. To solve this problem, the following strategies can be introduced: at the early stage of the oxygen blowing, the proposed control model is used to calculate the relative V and W and guide the BOF production. Then, the sublance technology is adopted at the late stage of oxygen blowing, because the physical and chemical reactions tend to be stable in this stage. Hence, the information of the melt liquid can be collected by the sublance. Therefore, another dynamic control model can be established with the sublance samples to achieve a higher end-point hit rate. For medium and small steel plants, the proposed static model is a suitable choice to reduce the consumption and save the cost.

Remark 3. Although this paper is mainly based on static control model, the prediction scheme is also compatible for other datasets over the globe. Especially, the proposed algorithm is competitive for the regression of time sequence datasets, such as the prediction of the blast furnace process and continuous casting process in the metallurgical industry.

5. Conclusion

In this paper, a WTWTSVR control model has been proposed. The new weighted matrix and the coefficient vector have been determined by the wavelet transform theory and added into the objective function of TSVR to improve the performance of the algorithm. The simulation results have shown that the proposed models are effective and feasible. The prediction error bound with 0.005% in C and 10°C in T can achieve a hit rate of 92% and 96%, respectively. In addition, the double hit rate of 90% is the best result by comparing with other three existing methods. The control error bound with 800 Nm^3 in V and 5.5 *tons* in W can achieve the hit rate of 90% and 88%, respectively. Therefore, the proposed method can provide a significant reference for real BOF applications. For the further work, on the basis of the proposed control model, a dynamic control model could be established to improve the end-point double hit rate of BOF up to 90% or higher.

Data Availability

The Excel data (Research Data.xls) used to support the findings of this study is included within the Supplementary Materials (available here).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by Liaoning Province PhD Start-Up Fund (No. 201601291) and Liaoning Province Ministry of Education Scientific Study Project (No. 2017LNQN11).

Supplementary Materials

The supplementary material contains the Excel research data for our simulations. There are 220 numbers of preprocessed samples collected from one steel plant in China. Due to the limitations of confidentiality agreement, we cannot provide the name of the company. The samples include the historical information of hot metal, total oxygen volume, total weight of auxiliary raw materials, and end-point information. In the excel table, the variables from column A to K are mainly used to establish the static control model for BOF. Columns H, I, J, and K are the output variables for C_model, T_model, V_model, and W_model in the manuscript, respectively. The input variables of the relative models are listed in Tables 1 and 2. By following the design procedure of the manuscript, the proposed algorithm and other existing algorithms can be evaluated by using the provided research data. *(Supplementary Materials)*

References

[1] C. Blanco and M. Díaz, "Model of Mixed Control for Carbon and Silicon in a Steel Converter," *Transactions of the Iron & Steel Institute of Japan*, vol. 33, pp. 757–763, 2007.

[2] I. J. Cox, R. W. Lewis, R. S. Ransing, H. Laszczewski, and G. Berni, "Application of neural computing in basic oxygen steelmaking," *Journal of Materials Processing Technology*, vol. 120, no. 1-3, pp. 310–315, 2002.

[3] A. M. Bigeev and V. V. Baitman, "Adapting a mathematical model of the end of the blow of a converter heat to existing conditions in the oxygen-converter shop at the Magnitogorsk Metallurgical Combine," *Metallurgist*, vol. 50, no. 9-10, pp. 469–472, 2006.

[4] M. Brämming, B. Björkman, and C. Samuelsson, "BOF Process Control and Slopping Prediction Based on Multivariate Data Analysis," *Steel Research International*, vol. 87, no. 3, pp. 301–310, 2016.

[5] Z. Wang, F. Xie, B. Wang et al., "The control and prediction of end-point phosphorus content during BOF steelmaking process," *Steel Research International*, vol. 85, no. 4, pp. 599–606, 2014.

[6] M. Han and C. Liu, "Endpoint prediction model for basic oxygen furnace steel-making based on membrane algorithm evolving extreme learning machine," *Applied Soft Computing*, vol. 19, pp. 430–437, 2014.

[7] X. Wang, M. Han, and J. Wang, "Applying input variables selection technique on input weighted support vector machine modeling for BOF endpoint prediction," *Engineering Applications of Artificial Intelligence*, vol. 23, no. 6, pp. 1012–1018, 2010.

[8] M. Han and Z. Cao, "An improved case-based reasoning method and its application in endpoint prediction of basic oxygen furnace," *Neurocomputing*, vol. 149, pp. 1245–1252, 2015.

[9] L. X. Kong, P. D. Hodgson, and D. C. Collinson, "Modelling the effect of carbon content on hot strength of steels using a

modified artificial neural network," *ISIJ International*, vol. 38, no. 10, pp. 1121–1129, 1998.

[10] A. M. F. Fileti, T. A. Pacianotto, and A. P. Cunha, "Neural modeling helps the BOS process to achieve aimed end-point conditions in liquid steel," *Engineering Applications of Artificial Intelligence*, vol. 19, no. 1, pp. 9–17, 2006.

[11] S. M. Xie, J. Tao, and T. Y. Chai, "BOF steelmaking endpoint control based on neural network," *Control Theory Applications*, vol. 20, no. 6, pp. 903–907, 2003.

[12] J. Jiménez, J. Mochón, J. S. de Ayala, and F. Obeso, "Blast furnace hot metal temperature prediction through neural networks-based models," *ISIJ International*, vol. 44, no. 3, pp. 573–580, 2004.

[13] C. Kubat, H. Taşkin, R. Artir, and A. Yilmaz, "Bofy-fuzzy logic control for the basic oxygen furnace (BOF)," *Robotics and Autonomous Systems*, vol. 49, no. 3-4, pp. 193–205, 2004.

[14] Jayadeva, R. Khemchandani, and S. Chandra, "Twin support vector machines for pattern classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 5, pp. 905–910, 2007.

[15] X. Peng, "TSVR: an efficient Twin Support Vector Machine for regression," *Neural Networks*, vol. 23, no. 3, pp. 365–372, 2010.

[16] D. Gupta, "Training primal K-nearest neighbor based weighted twin support vector regression via unconstrained convex minimization," *Applied Intelligence*, vol. 47, no. 3, pp. 962–991, 2017.

[17] R. Rastogi, P. Anand, and S. Chandra, "A ν-twin support vector machine based regression with automatic accuracy control," *Applied Intelligence*, vol. 46, pp. 1–14, 2016.

[18] M. Tanveer, K. Shubham, M. Aldhaifallah, and K. S. Nisar, "An efficient implicit regularized Lagrangian twin support vector regression," *Applied Intelligence*, vol. 44, no. 4, pp. 831–848, 2016.

[19] Y. Xu and L. Wang, "K-nearest neighbor-based weighted twin support vector regression," *Applied Intelligence*, vol. 41, no. 1, pp. 299–309, 2014.

[20] Y. Xu, X. Li, X. Pan, and Z. Yang, "Asymmetric v-twin support vector regression," *Neural Computing and Applications*, vol. no. 2, pp. 1–16, 2017.

[21] N. Parastalooi, A. Amiri, and P. Aliheidari, "Modified twin support vector regression," *Neurocomputing*, vol. 211, pp. 84–97, 2016.

[22] Y.-F. Ye, L. Bai, X.-Y. Hua, Y.-H. Shao, Z. Wang, and N.-Y. Deng, "Weighted Lagrange ε-twin support vector regression," *Neurocomputing*, vol. 197, pp. 53–68, 2016.

[23] C. Gao, M. Shen, and L. Wang, "End-point Prediction of BOF Steelmaking Based on Wavelet Transform Based Weighted TSVR," in *Proceedings of the 2018 37th Chinese Control Conference (CCC)*, pp. 3200–3204, Wuhan, July 2018.

[24] J. Shen and G. Strang, "Asymptotics of Daubechies filters, scaling functions, and wavelets," *Applied and Computational Harmonic Analysis* , vol. 5, no. 3, pp. 312–331, 1998.