A combined model of Hawkes-ICEEMDAN-Catboost-Holt for interval-valued carbon price forecasting

Muziyan Wang [1,*], Qingqing Zou [1], Qian Shi [1], Rui Luo [2]

[1] School of Business, Anhui University, Hefei, Anhui, China;

[2] School of Big Data & Statistics, Anhui University, Hefei, Anhui, China

* Corresponding Author Email: wangmuziyan216@163.com

Abstract. Accurate carbon price forecasting is crucial for efficiently operating both carbon trading and financial markets. This study proposes a Hawkes-ICEEMDAN-Catboost-Holt combined interval forecasting model to improve the accuracy of carbon price forecasts. First, the t-SNE-LLE second-order dimensionality reduction was performed on the collected search index data to quantify the attention of netizens. Subsequently, the attention series was converted into an interval series using the Hawkes exponential decay model, and further decomposed into radius and center series. In addition, the radius and center series were decomposed using ICEEMDAN and then reconstructed. Finally, the carbon price interval combined forecast was based on the Catboost–Holt model. The proposed model exhibits high accuracy and strong stability in comparison to other forecast models. The Hawkes exponential decay of netizens' attention is creatively included in the influence mechanism of carbon price fluctuations, effectively improving forecast accuracy. The interval decomposition and reconstruction methods provide an innovative development path for interval forecasting research. The residual and trend series obtained from interval reconstruction can be forecasted using Catboost and Holt, respectively, thus significantly improving the accuracy and stability of interval forecasts and enhancing the decision-making process for carbon trading and financial markets.

Keywords: Carbon price, Hawkes decay, combined forecast, Catboost, interval reconstruction.

1. Introduction

With the progress of global industrialization, the problem of greenhouse gas emissions has become increasingly serious. Excessive carbon emissions have caused global warming, which has become a major obstacle to sustainable development of the economy and society. Carbon dioxide produced by automobile exhaust and fossil fuels is one of the main causes of global climate problems. Therefore, controlling CO2 emissions has become a top priority, and countries worldwide are actively seeking countermeasures. As the 'world's largest greenhouse gas emitter, China is committed to controlling its carbon emissions to alleviate environmental pollution. To promote the low-carbon process, since 2013, the National Development and Reform Commission of China (NDRC) has established seven emission trading system (ETS) pilot projects, including Beijing, Tianjin, Shenzhen, Shanghai, Chongqing, Guangdong, and Hubei. In July 2021, the executive meeting of the State Council reviewed and approved the online trading of the national carbon emission rights trading market for the power generation industry and steadily expanded the coverage of industries to control and reduce greenhouse gas emissions through market mechanisms.

Economists consider carbon pricing one of the most efficient ways to reduce global warming [1]. At present, the forecast of carbon prices has become a research focus of Chinese scholars. Du et al. adopted a back propagation (BP) neural network model to study the mechanism of influencing factors of carbon prices from six different dimensions and forecast the carbon emission trading price well in Fujian, China [2]. Zhou et al. forecasted carbon prices using the model of long short-term memory network [3]. However, noting the existence of multiple prices of carbon within a day, it is difficult to forecast them accurately by using the characteristics of point-value time series. It is not difficult to find that most of the existing studies for carbon price forecast are based on a single mathematical model, and there are fewer studies on combined models. Ultimately, based on the existing research,

this study continues to advance by developing a combined Hawkes-ICEEMDAN-Catboost-Holt model for interval-valued carbon price forecast based on the public attention decay theory of Simkin et al [4], combined with unstructured data, and using the Baidu index, aiming to improve the accuracy of carbon price trend forecast.

2. Data source and preprocessing

The carbon price interval time series of the Hubei carbon market from January 2020 to December 2021 (https://www.hbets.cn/list/9.html) was selected as the sample set for analysis. The data from January 2020 to September 2021 were used as the training set, and the data from October 2021 to December 2021 were used as the prediction set.

The Baidu search index related to carbon prices was selected as the influencing factor for carbon prices in this study. The Baidu search index reflects the scale of the Baidu search and the changes in a keyword's upward and downward trends at a certain point of time or time period. Through the search index, we can further understand the exposure to related events and the attention of netizens to quantify the attention of keywords. Based on this study, 62 keywords were first selected and then combined with experts' suggestions and comprehensive relevance ranking, etc., for multiple rounds of screening, 19 keywords with high relevance to carbon trading prices, such as ERU, Carbon neutralization, and Global Climate Conference, were finally selected for multiple rounds of screening. The keyword search indexes from January 1, 2020, to December 31, 2021, were used to measure people's concerns about carbon trading. The data were then normalized, and dimensionality reduction was performed.

As carbon emissions trading prices are affected by weekends and holidays, there will be null values. The data have strong continuity since this study was based on daily search indices from 2020 to 2021. Therefore, this study adopts the direct deletion method for carbon price null values due to holidays, which directly deletes them and their related search indices on holidays.

3. Methodology

The general framework of this methodology is illustrated in Fig. 1. First, data were collected, and the carbon-price influence mechanism was constructed. The keywords affecting carbon prices were selected, and the time series of these influencing factors were extracted from the web pages and the carbon price series. Using t-SNE and LLE two-stage dimensional reduction to reduce them to 1 dimension, the sequence of public attention is constructed. Then, the daily decay values of the attention series were determined using the Hawkes exponential decay model to constitute the interval series of attention. In addition, the interval series of carbon price and attention are decomposed into the center and radius series, respectively. Subsequently, they are decomposed into IMFs using ICEEMDAN, and these IMFs are reconstructed into interval trends and residual series. Finally, the

residuals and trend series are predicted using CatBoost and Holt, respectively, and the results are integrated. A detailed description of this methodology is provided below.

3.1. Interval series construction of public attention

3.1.1 Dimensional reduction

Baidu Index data were used to reflect the level of public attention. Since the Baidu Index data collected are high-dimensional, we first reduced the data to 3-dimensional using t-SNE and visualized the data. Then, because LLE can learn the local linear low-dimensional manifold of arbitrary dimensions, and the algorithm is easy to implement, we used LLE to reduce the 3-dimensional data to 1-dimensional and get the attention sequence.

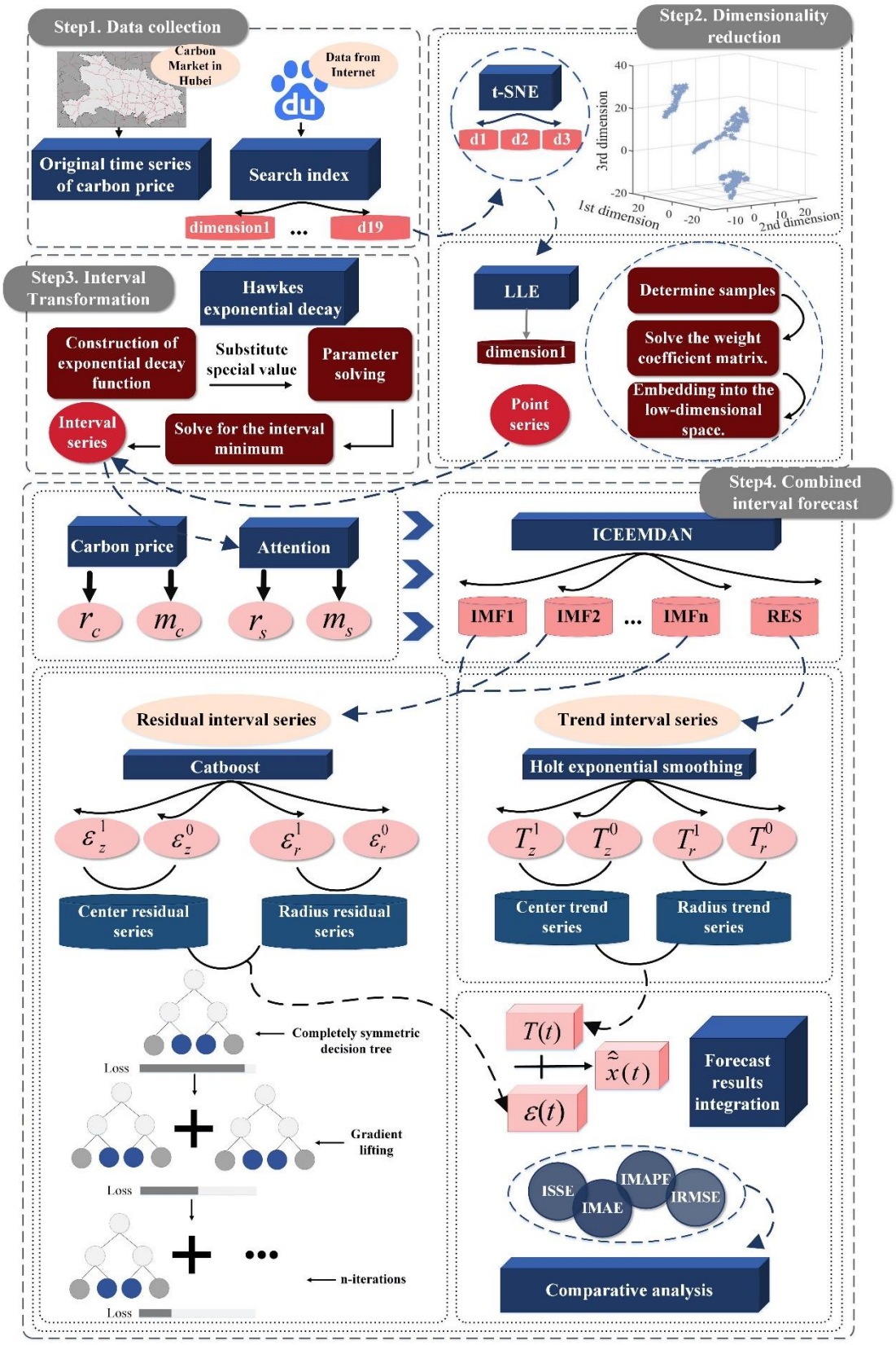


Figure 1. General frame diagram.

3.1.2 Hawkes exponential decay

Hawkes processes are also known as self/mutual-exciting processes [5], and the main idea is that historical events have a motivating effect (positive effect) on future events, and that the effect of historical events on the future is assumed to be monotonically exponentially decreasing and then superimposed in a cumulative form. Referring to Simkin's study on the exponential decay of online

news influence and the law of radioactive decay in physics, this paper assumes that public attention decays exponentially over time. For event $u \in C$ in the U dimensional Hawkes process, the conditional intensity is expressed as $\lambda(t) = \mu(t) + \sum_{u \in C} \sum_{t_i < t} k(t - t_i)$ where $\mu(t)$ denotes the basic attention of the keyword; k is a kernel function, and F denotes the time decay effect of the keyword search index, which is commonly used in the form of an exponential kernel function with the formula: $k(t) = \alpha e^{-\beta(t)}$ where α denotes the keyword attention and β denotes the control decay rate.

The parameters in the Hawkes process are generally solved by optimizing the log-likelihood function. Assume that $A = (\alpha) \in R^{U \times U}$ is the influence matrix between search indices. For the ensemble of search index sequence $S = \{S_1, S_2, \cdots S_m\}$, each search index sequence $S_i = \{(a_{ij}, t_{ij})\}_{j=1}^{n_i}$ is the ensemble of search indices for the time interval $[0, T_i]$, where each (a_{ij}, t_{ij}) denotes the time of occurrence of the search index a_{ij} as t_{ij}. The log likelihood corresponding to the parameters of the Hawkes process model $\Theta = \{\mu, A\}$ is: $\gamma(\mu, A) = \sum_{i=1}^{m} \left(\sum_{j=1}^{n_i} \log \lambda_{a_{ij}}(t_{ij}) - \sum_{u=1}^{U} \int_0^{T_i} \lambda_u(t) dt \right)$. Since the influence matrix A is usually sparse and of low rank, the matrix A is restricted here by adding Lasso and kernel parametrization as penalty terms, which are calculated as: $\min_{\mu \geq 0, A \geq 0} -\gamma(\mu, A) + \rho_1 \|A\|_1 + \rho_2 \|A\|_*$ where $\|\cdot\|_1$ denotes the L_1 norm, $\|\cdot\|_* = \sum_{i=1}^{rank} \sigma_i$ denotes the kernel norm, and the parameters ρ_1 and ρ_2 control the weights of the different penalty terms.

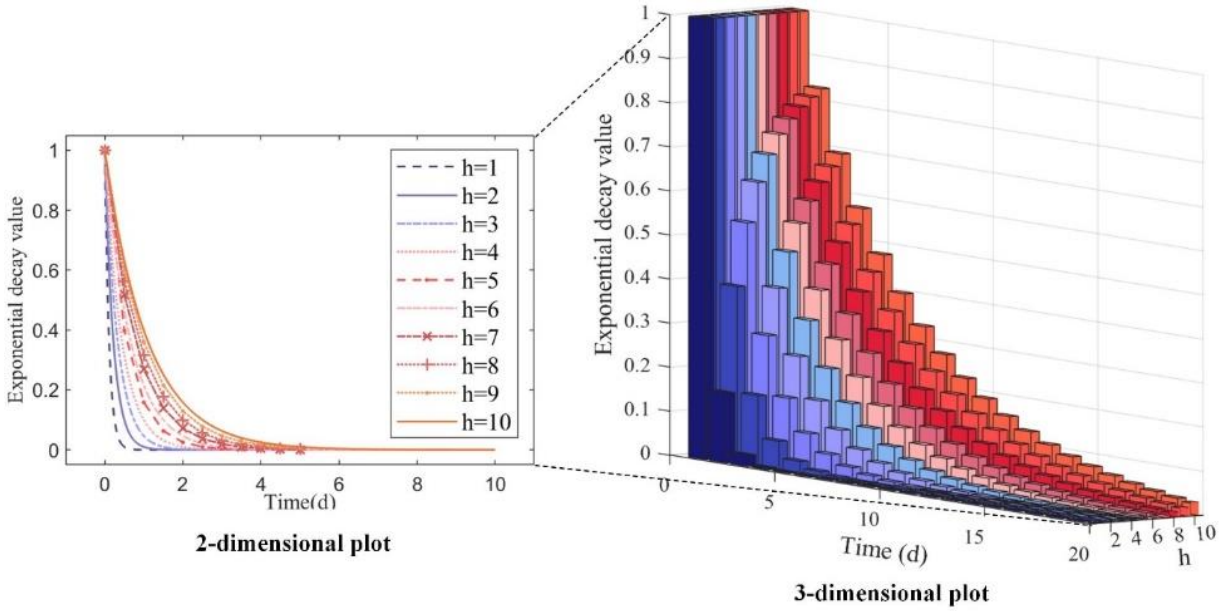


Figure 2. Hawkes exponential decay.

To solve the decay rate β, we assume that the time interval when the search index attention decays to 0.0001 is the endpoint t_{end}. That is, if the search index attention decays to 0.0001 on days h, $t_{end} = h$, then the decay rate $\beta = \frac{-\ln(0.0001/\alpha)}{h}$ can be further calculated. Different decay rates β can be derived according to the different decay days in h. From Fig. 2, it can be observed that a smaller h causes a faster decay of the index curve. The correlation between the decay series and carbon price series under different h values can be compared, and the best decay days are obtained when the correlation is the largest. Finally, the optimal decay rate was determined.

By solving these parameters, we can derive the attention of the Baidu index at the moment of t after decaying for one day, which is denoted as $\mu'(t)$. In this study, $\mu'(t)$ is regarded as the lowest value of attention of the Baidu index in a day, whereas the basic intensity $\mu(t)$ is regarded as the highest value of attention, thus constituting the interval of attention of the Baidu index on that day $[\mu'(t), \mu(t)]$.

3.2. Combined interval forecast method

The model framework of the proposed carbon price-combined interval forecasting method is shown in Fig. 1, with the details of the specific steps as follows:

Step 1 Interval decomposition. The attention interval series and carbon price time series obtained from the previous quantification and intervalization were decomposed separately to obtain the attention center series m_s, attention radius series r_s, carbon price center series m_c and carbon price radius series r_c.

Step 2 ICEEMDAN decomposition. ICEEMDAN decomposition is performed for ms, rs, mc, and r_c, respectively, so that they are decomposed into multiple eigenmode functions IMFs and res trend terms.

Step 3 Interval reconstruction. The IMFs and res are reconstructed into trend series and residual series, which are divided into attention radius trend series T_r^1, attention radius residual series ε_r^1, attention center trend series T_z^1, attention center residual series ε_z^1; carbon price radius trend series T_r^0, carbon price radius residual series ε_r^0, carbon price center trend series T_z^0, and carbon price center residual series ε_z^0. Then, the radius residual series and center residual series were reconstructed into the residual interval series $\varepsilon(t)$, and the radius trend series and center trend series were reconstructed into the trend interval series $T(t)$.

Step 4 Single-item forecasting. CatBoost and Holt two-parameter models were used for forecasting the residual series and trend series, respectively. The attention series was used as the eigenvalue input for the carbon price time-series prediction.

Step 5 Reconstruct again. The interval trend and residual predictions were summed to obtain the predicted values of the interval time-series $x(t) = [x_l(t), x_h(t)] = [T_l(t) + \varepsilon_l(t), T_h(t) + \varepsilon_h(t)]$.

Step 6 Error evaluation. Error evaluation indexes, such as IMAE, LSSE, and IMAPE, were used to compare and analyze the prediction accuracy of the model in this study with other advanced models.

3.3. Forecast method and evaluation metrics

(1) CatBoost is an improved algorithm based on the GBDT model proposed by Prokhorenkova et al. [6] with a symmetric decision tree as the base model, which has a very high forecast efficiency. The CatBoost model has fewer parameters and high accuracy and embeds an innovative algorithm that automatically processes category-based features into numerical features. Moreover, the model addresses the problems of gradient bias and forecast shift, thus reducing the need for hyperparameter tuning, avoiding overfitting, greatly improving the accuracy of the algorithm, and possessing great robustness. The model consists of two parts: categorical features and ordered boosting.

(2) The Holt two-parameter exponential smoothing method. It is an exponential smoothing model by Charles C [7]. Holt that extends the linear trend component of the primary exponential smoothing method to provide more accurate forecasts of trending data.

(3) Error evaluation index. Based on the error evaluation index of the point forecast, this study defines the error evaluation index used to measure the forecast results, as shown in Table 1.

Table 1. Evaluation metrics.

Metric	Definition	Equation
ISSE	Interval sum of squares due to error	$ISSE = \frac{1}{2}\sum_{t=1}^{n}[(x_l(t)-x_l(t))^2+(x_h(t)-x_h(t))^2]$
IMAE	Interval mean absolute error	$IMAE = \frac{1}{2n}\sum_{t=1}^{n}[\lvert x_l(t)-x_l(t)\rvert+\lvert x_h(t)-x_h(t)\rvert]$
IMAPE	Interval mean absolute percentage error	$IMAPE = \frac{1}{2n}\sum_{t=1}^{n}[\lvert\frac{x_l(t)-x_l(t)}{x_l(t)}\rvert+\lvert\frac{x_h(t)-x_h(t)}{x_h(t)}\rvert]$
IRMSE	Interval root mean square error	$IRMSE = \frac{1}{2n}[\sqrt{\sum_{t=1}^{n}(x_l(t)-x_l(t))^2}+\sqrt{\sum_{t=1}^{n}(x_h(t)-x_h(t))^2}]$

4. Empirical study and analysis

4.1. Decomposition and reconstruction of interval series

Based on the nonlinearity and non-smoothness of the carbon price interval time series, the center series $m(t)$ and radius series $r(t)$ of both the carbon price and search index interval time series were separately decomposed using the ICEEMDAN decomposition method. Fig. 3 displays part of the original time series along with their decomposition and reconstruction results. Specifically, using the carbon price time series as an example, the figure shows the decomposition results of the center series $m(t)$. The IMF1, IMF2,..., IMF6 are the decomposed fluctuation series of $m(t)$, while 'res' denotes the trend term series. Fig. 3 presents the decomposition results of the radius series $r(t)$, where the IMF1, IMF2,..., and IMF7 are the decomposed fluctuation series of $r(t)$ and 'res' denotes the trend term series.

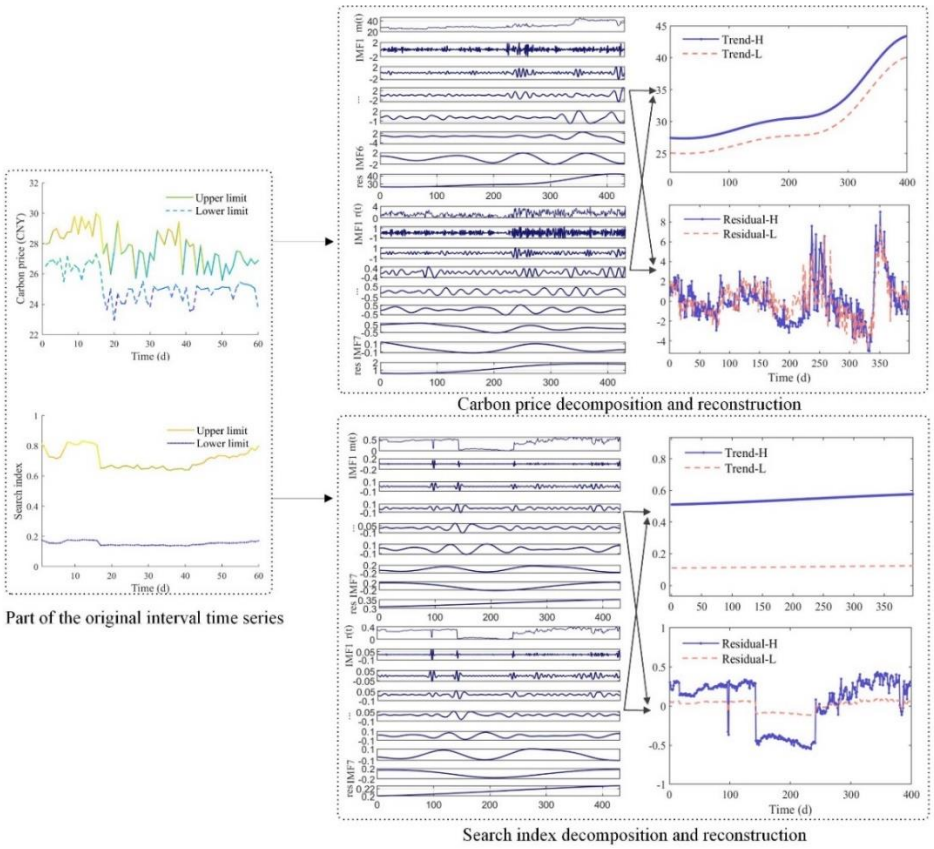


Figure 3. Decomposition and reconstruction.

The interval trend series T and residual series ε are then obtained by interval reconstruction. First, the residual series $\varepsilon_m = \sum_{i=1}^{6} IMF_{mi}$ of the center $m(t)$ and the trend series $T_m = res_m$ of the center $m(t)$ are obtained. Then, the residual series $\varepsilon_r = \sum_{i=1}^{7} IMF_{ri}$ of the radius $r(t)$ and the trend series $T_r = res_r$ of the radius $r(t)$ are obtained.

The trend series of the interval time series $T(t) = [T_l(t), T_h(t)] = [T_m - T_r,\ T_m + T_r]$ and the residual series of the interval time series $\varepsilon(t) = [\varepsilon_l(t), \varepsilon_h(t)] = [\varepsilon_m - \varepsilon_r, \varepsilon_m + \varepsilon_r]$ are obtained by interval reconstruction, as shown in Fig. 3.

4.2. Results and discussion

The decomposed and reconstructed interval trend series $\tilde{T}(t)$ and interval residual series $\tilde{\varepsilon}(t)$ are forecasted using Holt's two-parameter exponential smoothing method and CatBoost, respectively, to obtain the forecast interval trend series $\hat{\tilde{T}}(t) = [\hat{T}_l(t), \hat{T}_h(t)]$ and forecast interval residual series $\hat{\tilde{\varepsilon}}(t) = [\hat{\varepsilon}_l(t), \hat{\varepsilon}_h(t)]$. The final forecast interval series $x = [\hat{T}_l(t) + \varepsilon_l(t), \hat{T}_h(t) + \varepsilon_h(t)]$ of the carbon price was obtained by reconstructing, as shown in Fig. 4.

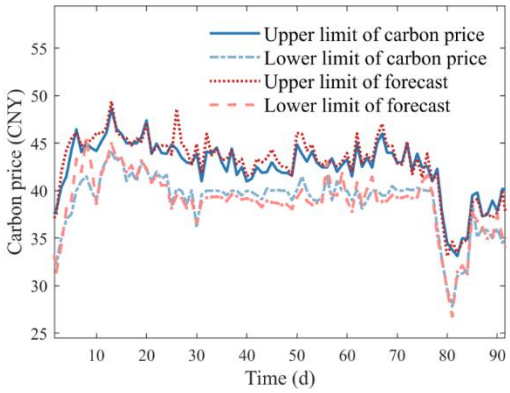


Figure 4. Forecast results of carbon price.

To verify the forecast effect of the model proposed in this study, the model is compared and analyzed based on four error evaluation indicators and comparison experiments I-V.

Experiment I: Comparison between classic individual models.

The five classic individual forecasting models Holt, CatBoost, LSTM, Bi-LSTM, and Bootstrap (Model 1~5) are compared. In order to control variables, the five individual models in the are all based on the Hawkes decay and combined forecast model proposed in this paper. Table 2 and Fig. 5 display that Model 2 performed better among the five individual forecasting models, and its four error indicators were the smallest among the five individual forecasting models, indicating that CatBoost has an advantage in reducing carbon price forecasting errors. The error of Model 3 is extremely large, in which the ISSE reaches 10915.96, which is 25 times higher than that of Model 8. The IMAPE of Model 3 is 24 times higher than that of Model 6, and the IMAE and IRMSE of Model 3 are also much higher than those of other individual forecasting models and Model 6, which indicates that LSTM is not applicable to the prediction of carbon trading price series in this study.

Experiment II: Comparison of combined model and individual models.

Comparing the individual forecast models Holt and Catboost (Model 1~2) with the combined forecast model Catboost-Holt (Model 8) in this paper, the prediction error of the combined forecast model (Model 8) is significantly smaller than that of the individual forecast models (Model 1~2), indicating that the combined forecast model can extract multi-scale data features, further improving the prediction accuracy and having a better fitting effect.

Experiment III: Comparison with model without Hawkes exponential decay.

As shown in Table 2 and Fig. 5, experiment III compares the proposed combined forecast model without Hawkes exponential decay (Model 6) with the textual model (Model 8). It shows that the error of Model 8 is significantly smaller than that of Model 6, indicating that the decay of public attention provides effective information for carbon price forecasts.

Experiment IV: Comparison with point-valued forecasting method.

The carbon price point-valued combined forecast model (Model 7) is compared with the interval-valued combined forecast model (Model 8). Model 7 is a point-valued forecast model based on the combined model used in this study. Table 2 and Fig. 5 show that the error of the interval-valued combined forecast model is significantly smaller than that of the point-valued combined forecast model, indicating that interval-valued forecasting is more accurate and stable than point-valued forecasting.

Table 2. Error evaluation index analysis of comparison experiments I-IV.

No.	Forecast model	ISSE	IMAPE	IMAE	IRMSE
1	Holt	4239.64	0.13	7.45	3.87
2	Catboost	732.78	0.02	6.74	2.78
3	LSTM	6252.25	0.24	11.56	5.96
4	Bi-LSTM	5344.33	0.34	8.13	4.33
5	Bootstrap	3963.98	0.17	7.78	3.98
6	Model without Hawkes	624.49	0.26	5.84	2.49
7	Point-valued forecast	574.34	0.18	7.47	3.26
8	**Textual model**	**251.97**	**0.01**	**3.94**	**1.23**

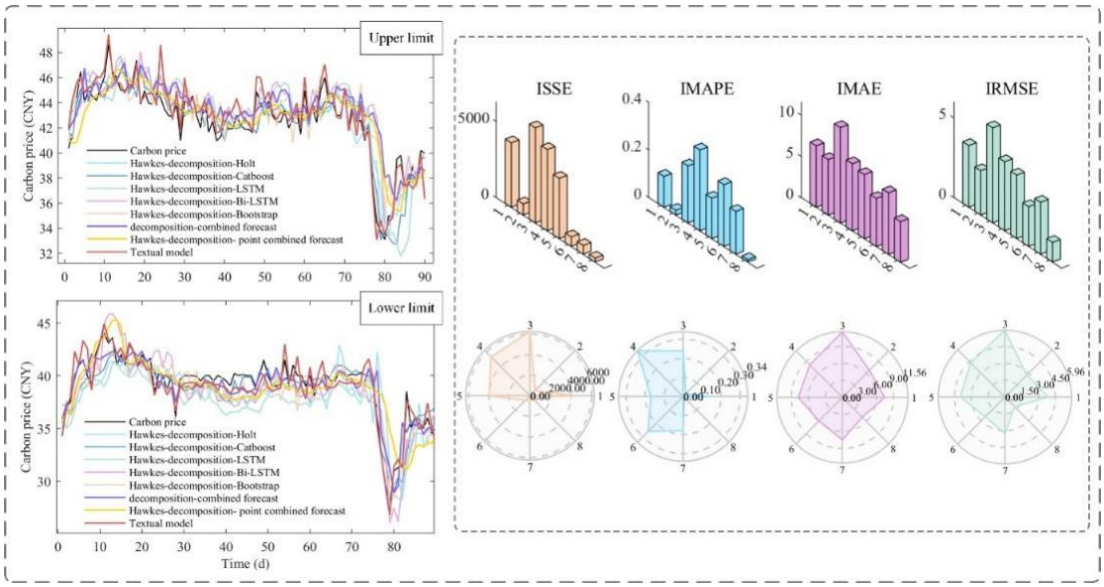


Figure 5. Error plots and prediction results of comparison experiments I-IV.

Experiment V: Comparison with different interval decomposition and recons-truction methods.

Experiment V aims to compare this model with the recent decomposition methods. For example, Sun W [8] decomposed the carbon price series using Wavelet; Ji Z [9] successfully predicted carbon price based on decomposition-integration method; Zhang J L [10] made interval prediction of electricity price based on IEMD as well as ARMAX-EGARCH-ANFIS model; Zhou F [11] used CEEMDAN-LSTM combined model to predict carbon price and achieved high prediction accuracy.

Experiment V is designed to compare different interval decomposition and reconstruction methods, including Wavelet, SSA, IEMD, CEEMDAN, combined model without decomposition and the

textual model (Model 5). In order to control variables, the five decomposition and reconstruction models are all based on the Hawkes decay and combined forecast model proposed in this paper.

Table 3. Error evaluation index analysis of comparison experiments V

No.	Forecast model	ISSE	IMAPE	IMAE	IRMSE
1	Wavelet	379.64	0.16	5.03	3.07
2	SSA	285.26	0.08	4.03	1.98
3	IEMD	332.85	0.11	4.35	2.67
4	CEEMDAN	319.69	0.09	4.01	1.87
5	combined model without decomposition	475.34	0.31	5.67	3.34
6	**Textual model**	**251.97**	**0.01**	**3.94**	**1.23**

As shown in Table 3 and Fig. 6, it is easy to find that the interval series decomposition model proposed in this paper has the optimal effect. In Table 3, ordinary time series decomposition models such as Wavelet, SSA and CEEMDAN have ISSE, IMAE, IMAE and IRMSE values around 300, 0.1, 4 and 2, among which Wavelet performs the worst in the evaluation index, which indicates that under the same conditions, Wavelet is not able to decompose the carbon price series and attention series with great fluctuation and nonlinear scientifically. According to Table 3 and Fig. 6, we can also find IEMD and the textual decomposition model has a smaller error than Wavelet, SSA and CEEMDAN. This indicates that the interval decomposition model is significantly better than the ordinary decomposition model in terms of error evaluation indicators. This shows that since carbon price has multiple prices in one day, such as opening price, closing price, highest price and lowest price, and has obvious interval series characteristics, interval decomposition is more suitable for studying carbon price fluctuation.

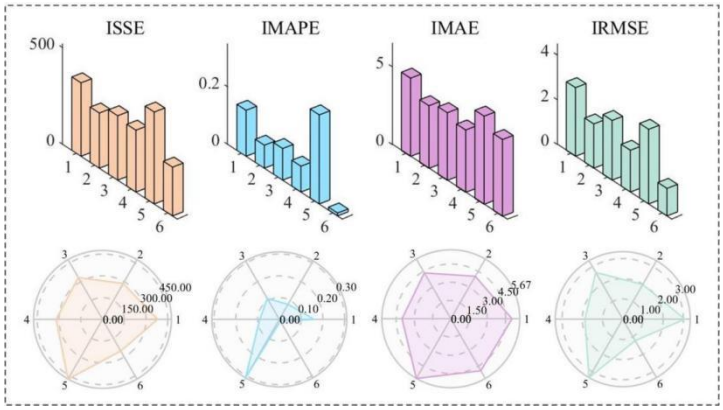


Figure 6. Error plots of comparison experiments V.

Table 3 and Fig. 6 also show that the error of the textual model (Model 6) is significantly smaller than that of the combined model without decomposition and reconstruction (Model 5), indicating that the ICEEMDAN decomposition and interval reconstruction methods can fully extract the characteristics of the interval series.

In summary, it is found that the combined forecast has certain advantages over a single forecast and we combine the single models with good prediction effect. Application of the Hawkes exponential decay model can reduce the forecast error. Interval-valued forecasts can extract nonlinear information more fully and improve forecast accuracy compared to point-valued forecasts. The proposed decomposition and interval reconstruction of data and the combination of different forecast models for trend series and residual series can also effectively improve the forecast accuracy. The experimental results show that the combined forecast model constructed in this study has better

accuracy, superiority, and applicability, absorbs decay information, and proposes an advanced processing method for interval construction.

5. Conclusion

In this study, the Hawkes exponential decay model was applied to convert the attention sequence into an interval series and an advanced method for interval decomposition and reconstruction of interval-valued series was proposed. The residual series and trend series were forecasted using the CatBoost and Holt models, respectively, and the forecast results were integrated. Finally, the validity of the model was verified against the results of an empirical study and other comparison models. Based on this study, the following conclusions were drawn.

(1) The Hawkes exponential decay can effectively measure the decay of public attention. A sequence of attention intervals was constructed by calculating the daily decay in public attention. Introducing the public attention interval series into the forecasting of carbon prices significantly improved their impact.

(2) The novel decomposition and interval reconstruction method converts the interval-valued data into interval trends and residuals. It can adequately extract the features hidden in the original interval time series.

(3) The Holt two-parameter exponential smoothing and CatBoost were used to forecast the trend and residual series for the features of the different modal data, respectively. The combination of these two excellent methods yields satisfactory results.

(4) The designed combined interval-valued forecasting model based on Hawkes decay is highly accurate and applicable, which is better than the comparison models mentioned above. It provides an efficient technical tool for the healthy development of the carbon market.

References

[1] Chevallier J. Nonparametric modeling of carbon prices[J]. Energy Economics, 2011, 33(6): 1267-1282.

[2] Du Y, Chen K, Chen S, et al. Prediction of Carbon Emissions Trading Price in Fujian Province: Based on BP Neural Network Model[J]. Frontiers in Energy Research, 2022: 1049.

[3] Zhou F, Huang Z, Zhang C. Carbon price forecasting based on CEEMDAN and LSTM[J]. Applied energy, 2022, 311: 118601.

[4] Simkin M V, Roychowdhury V P. Why Does Attention to Web Articles Fall with Time? [J]. Journal of the Association for Information Science & Technology, 2015, 66: 1847-1856.

[5] Xiao DING, Jihao SHI, Junwen DUAN, Bing QIN, Ting LIU. Quantifying the effects of long-term news on stock markets on the basis of the multikernel Hawkes process[J]. Science China (Information Sciences), 2021, 64(09): 44-56

[6] Prokhorenkova L, Gusev G, Vorobev A, et al. CatBoost: unbiased boosting with categorical features[J]. Advances in neural information processing systems, 2018, 31：1-9

[7] Liu L, Wu L. Predicting housing prices in China based on modified Holt's exponential smoothing incorporating whale optimization algorithm[J]. Socio-Economic Planning Sciences, 2020, 72: 100916.

[8] Sun W, Xu C. Carbon price prediction based on modified wavelet least square support vector machine[J]. Science of the Total Environment, 2021, 754: 142052.

[9] Ji Z, Niu D, Li M, et al. A three-stage framework for vertical carbon price interval forecast based on decomposition–integration method[J]. Applied Soft Computing, 2022, 116: 108204.

[10] Zhang J L, Zhang Y J, Li D Z, et al. Forecasting day-ahead electricity prices using a new integrated model[J]. International journal of electrical power & energy systems, 2019, 105: 541-548.

[11] Zhou F, Huang Z, Zhang C. Carbon price forecasting based on CEEMDAN and LSTM[J]. Applied energy, 2022, 311: 118601.