Space-time measurement model of carbon emission based on deep learning

Keming Cheng[1,a,*], Xinshi Ou[1,b], Yangdi Hu[1,c]

[1]School of Finance, Southwestern University of Finance and Economics, Chengdu, China

[a] 2532809313@qq.com, [b]ouxinshi500@qq.com, [c]1125276233@qq.com

*corresponding author

Keywords: carbon emission, Space-time model, TOPSIS legal right Act, SVM model, F model, ARIMA model, LSTM

Abstract: As the global temperature rises, the greenhouse effect becomes more and more obvious, reducing carbon dioxide emissions has become a topic of global concern. Through the background study of carbon emissions, this paper first conducts regional research based on carbon emissions data of China carbon accounting database. Secondly, based on the data of the China Urban Statistical Yearbook, this paper studies the carbon emissions of various regions in the country from 1997 to 2020, explores the spatio-temporal evolution characteristics of each region, and analyzes its influencing factors by constructing the "big data" element momentum entry point. In order to eliminate the error factor as much as possible, a dimension reduction model based on grey correlation analysis was first established to eliminate the factors that have a small impact on carbon emissions. Then, the topsis score comprehensive evaluation model based on entropy weight method is constructed, which shows that carbon emission level is closely related to economy and policy. Finally, SF model, LSTM model, ARIMA time prediction model and SVM model based on factor analysis and genetic algorithm are established to model data indicators, and the optimal model is obtained by comparing the models. Finally, some suggestions on carbon emission are given.

1. Introduction

For a long time, carbon dioxide emissions from fossil fuels are the main cause of the greenhouse effect and global warming. In 2007, China ranked first in global carbon emissions, becoming a "carbon emission" country. The development of regional economy is closely related to the quality of national economy, and the promotion of urban industrialization leads to carbon emissions, which has become the focus of attention. As the main contradiction in Chinese society shifts from development to meeting the needs of the people for a better life, uneven regional development has led to alienation and imbalance in carbon emissions.

Tan Dan et al. (2008) [1] conducted a study on carbon emissions related to global warming, and analyzed the total amount, characteristics and differences of carbon emissions in the eastern, central and western regions of China, as well as their correlation with GDP. Wang Feng et al. (2010) [2] studied the driving factors of carbon emission growth in China's economic development and summarized relevant research results. Yang Sai et al. (2012) [3] used inter-provincial panel data to explore regional differences in carbon emissions and summarized the research results. Yuan Yuan et al. (2016) [4] studied the impact of industrial structure on carbon emission through multi-country data. Chen Zhanming et al. (2018) [5] extended the STIRPAT model based on the panel data of cities above prefecture level in China to study the impact of urban factors on CO2 emissions. He Wenju et al. (2019) [6] analyzed the spatial aggregation effect of carbon emissions in the context of urbanization in China and put forward relevant suggestions. Du Haibo et al. (2021) [7] studied the spatial-temporal pattern evolution and influencing factors of carbon emissions from energy consumption in the Yellow River Basin. Deng Rongrong et al. (2021) [8] studied the path of carbon peaking and carbon neutrality in the era of digital economy, and reduced carbon emission intensity with the help of economic growth, industrial structure and technological innovation effects. These studies deeply expounded the status quo, influencing factors and future development trends of China's

  DOI: 10.25236/icceme.2024.002

carbon emission research, and provided policy recommendations for promoting green and low-carbon development and solving challenges.

2. Model Establishment and Solution

2.1. Data Preprocessing

Standardization is generally the mapping of data to a specified range, which is used to remove dimensions and dimensional units of data of different dimensions. Common mapping ranges are [0,1] and [-1,1]. This paper adopts Min-Max normalization, also known as deviation normalization, which is a linear transformation of the original data, so that the resulting value is mapped between [0,1]. The conversion function is as follows:

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

This normalization method is more suitable for the case of relatively concentrated values. At the same time, KMO and Bartlett validity analyses were also carried out on the data in this paper, and the results are shown in Table 1:

Table 1: KMO and Bartlett's Checklist

KMO Value		0.72
Bartlett sphericity test	Approximate chi-square	1084.929
	df	1
	P	4.15E-10

2.2. Topsis Comprehensive Evaluation Index Based on Entropy Weight Method

Grey Relation Analysis (GRA) is a method used to study the correlation between variables, especially when sample data is small, sample features are missing, or data quality is not high. It is developed from grey system theory, which aims to analyze and describe the degree of correlation between variables.

Through the calculation of grey correlation degree, the correlation between each variable can be obtained, and then the data analysis and decision support can be carried out.

In this paper, based on grey correlation analysis, carbon emissions of prefecture-level cities are studied and analyzed in depth. This paper collects carbonite energy reserves, energy prices, carbon emissions of various industries and so on. At the same time, in order to simplify the model and remove the influence of data noise, this paper establishes a dimension reduction model based on gray correlation analysis, and selects nine indicators including energy reserve, energy price ratio (prefecture-level city/national average) and population density. In order to better equal rights, a comprehensive evaluation model of topsis score based on entropy weight method is established, and the results are shown in Figure 1 below.

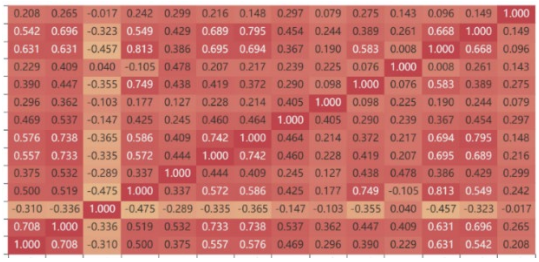

Figure 1: Grey correlation analysis results

In order to make the measurement more scientific, we need to give the corresponding weights of each index. In this paper, entropy method is used to assign weights.

Set up the index system has y vintage, n evaluation objects, m evaluation index, α for years ($\alpha = 1, 2, \ldots, y$), i is the evaluation object ($i = 1,2,\ldots,n$), j is the evaluation index ($j = 1,2,\ldots,m$), $x_{\alpha_{ij}}$

is the standardized value of the j evaluation index of the i evaluation object in the α year. It can be divided into the following three parts:

Calculate the proportion of each index $P_{\alpha_{ij}}$:

$$P_{\alpha_{ij}} = \frac{x_{\alpha_{ij}}}{\sum_{\alpha=1}^{y}\sum_{i=1}^{n} x_{\alpha_{ij}}} \tag{2}$$

Calculate the entropy e_i and difference coefficient g_j of each index:

$$e = -k \sum_{\alpha=1}^{y\sum g_j} \sum_{i=1}^{n\sum \alpha_{ij}} p_{\alpha_{ij}} ln \tag{3}$$

Calculate the weight of each indicator w_j:

$$w_j = \frac{g_j}{\sum_{j=1}^{m} g_j} = \frac{1 - e_j}{m - \sum_{j=1}^{m} e_j} \tag{4}$$

Calculate positive and negative ideal solutions:

$$x_j^+ = \{(max(x_{\alpha ij}|j \in J)|1 \le \alpha \le y, 1 \le i \le n\},\ x_j^- = \{(min(x_{\alpha ij}|j \in J)|1 \le \alpha \le y, 1 \le i \le n\} \tag{5}$$

Calculate the distance d_i^+ of the evaluation object to the positive ideal solution x^+ and the Euclidean distance d_i^- of the negative ideal solution x^-

$$d_i^+ = \sqrt{\sum_{a=1}^{y}\sum_{j=1}^{m}(w_j(x_{aij} - x_j^+))^2}, \qquad d_i^- = \sqrt{\sum_{a=1}^{y}\sum_{j=1}^{m}(w_j(x_{aij} - x_j^-))^2} \tag{6}$$

Calculate the relative progress of each evaluation object C_i^*

$$C_i^* = \frac{D_i^-}{D_i^+ + D_i^-}, i = 1,2,\dots,n \tag{7}$$

2.3. Random Forest (RF) model

Random forest model is a classic Bagging model, and its weak learner is decision tree model. The random forest model will randomly sample the original data set to form n different sample data sets, and then build n different decision tree models according to these data sets, and finally obtain the final result according to the average value or voting situation of these decision tree models.

The random forest model used in this paper combines multiple classification trees through self-sampling for voting classification. In the establishment of decision tree, complete splitting is adopted without pruning, because random sampling ensures randomness and avoids overfitting.

The importance of random forest results can be seen that the policy influence coefficient is larger, followed by the three wastes emission. The predicted results are shown in Figure 2 and Table 2.

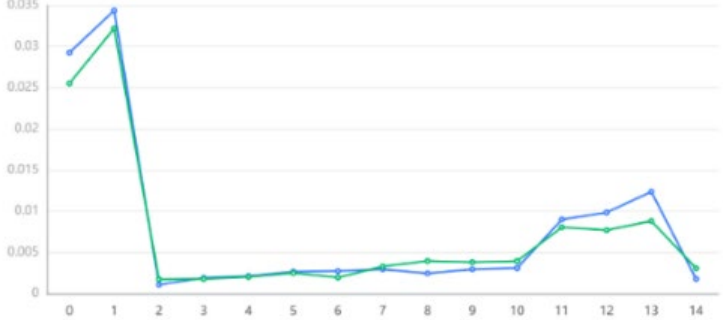

Figure 2: Random forest prediction results

Table 2: Random Forest Regression Results

	MSE	RMSE	MAE	MAPE	R^2
Training set	0.062	0.44133745	0.651	0.194	0.876
Cross verification set	0.078	0.49501942	0.676	0.291	0.852
Test set	0.085	0.51675465	0.721	0.302	0.765

2.4. Long Term Memory Network Model (LSTM)

Compared with the traditional recurrent neural network model, LSTM introduces three gates (input gate, forgetting gate, output gate) and one cell state. These mechanisms enable LSTM to better deal with long-term dependencies in the sequence. In order to establish a suitable LSTM model, this paper starts from the structure of the model.

The forget is the following equality relations:

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{8}$$

The input gate can be written as:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i), \tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C), C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{9}$$

The expression of output gate is:

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{10}$$

Then, take C_t as the input to the tanh function, determine the C_t part of the final output, and multiply it with o_t to get the final output, i.e:

$$h_t = o_t * \tanh(C_t) \tag{11}$$

The result is shown in Figure 3:

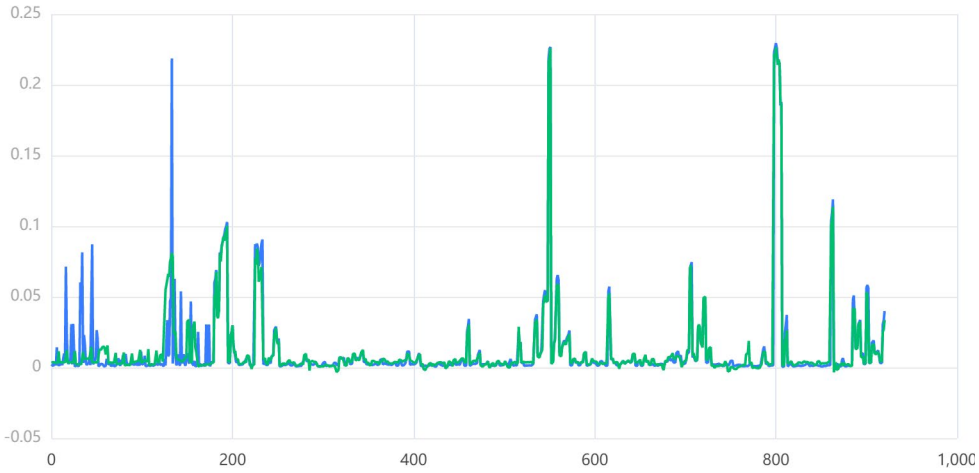

Figure 3: LSTM model prediction results

2.5. Auto-Regressive Moving Average (ARIMA) Model

The ARIMA model consists of three parts, namely, autoregression (AR), difference (I) and moving average (MA). Their meanings are as follows: AR represents autoregression; Describe the relationship between the current value and the historical value, and use the historical time data of the variable itself to predict itself.

By using this model, the prediction random fluctuation in the original AR model can be eliminated well. To improve the accuracy of the model; I stands for single integral order. By combining the autoregressive model (AR), moving average model (MA) and difference method (I), we get the differential autoregressive moving average model ARIMA(p, d, q), where d is the order of data difference, and ARIMA is the ARMA model after difference. The test table obtained is shown in Table 3:

Table 3: ADF Test List

	Difference order	t	P	AIC	Threshold		
					1%	5%	10%
CO_2	0	-7.254	0.000***	0.619	1.431	2.862	2.567
	1	-27.531	0.000***	0.327	1.431	2.862	2.567
	2	-33.321	0.000***	0.19	1.431	2.862	2.567

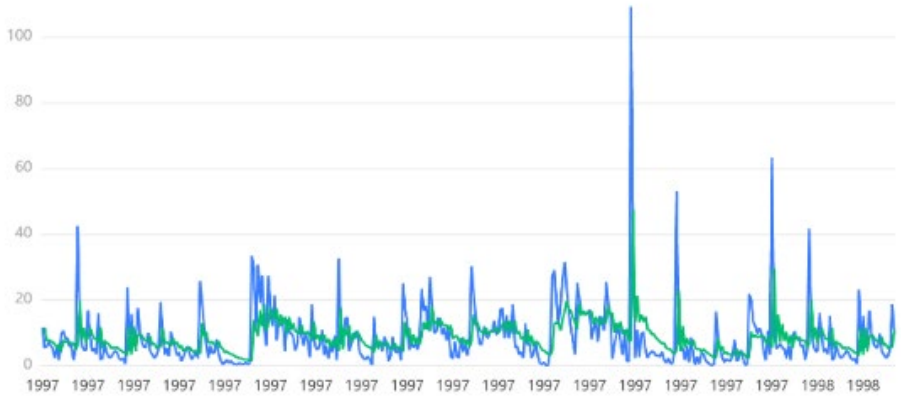

Figure 4: ARIMA Model fitting renderings (1997-1998)

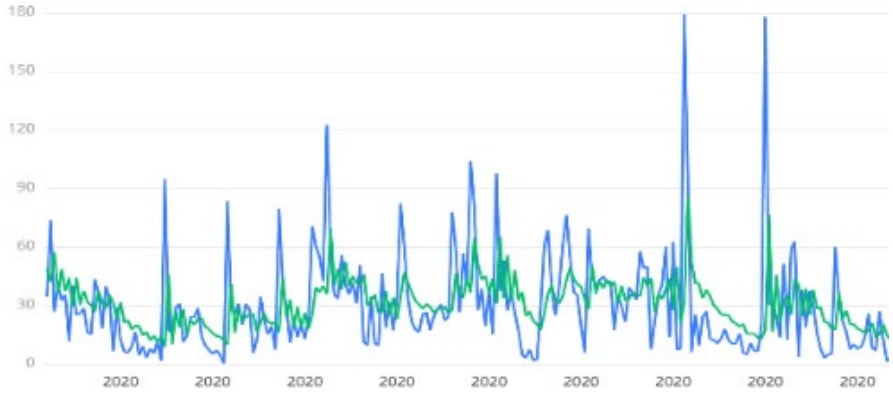

Figure 5: ARIMA Model fitting renderings (2020-)

From figure 4 and figure 5, it is found that the fitting of the overall trend of the model is more appropriate, but for some peak points, the fitting effect of the maximum and minimum values is worse, and the actual value may be 2-3 times of the predicted value.

2.6. Support Vector Machines (SVM) model

This paper assumes that the training set on the given feature space is:

$$T = \{[x_1, y_1], [x_2, y_2], \cdots [x_n, y_n]\} \tag{12}$$

For the given data set T and the hyperplane $wx + b = 0$, the geometric interval of the hyperplane with respect to the sample point (x_i, y_i) can be obtained as:

$$\gamma_i = y_i \left(\frac{w}{||w||} \cdot x_i + \frac{b}{||w||} \right) \tag{13}$$

Note the minimum value of the geometric interval of all sample points on the hyperplane as:

$$\gamma = \min_{i=1,2,\cdots,N} \gamma_i \tag{14}$$

The SVM model solving problem constructed in this paper can be transformed into the optimal value solving problem under the following specified constraints, namely:

$$\max_{w,\ b} \gamma$$

$$s.t. \quad y_i\left(\frac{w}{||w||}\cdot x_i + \frac{b}{||w||}\right) \geq \gamma, i = 1,2,\cdots,N \tag{15}$$

This paper uses hinge loss, so the original problem can be rewritten as

$$\min_{w,\ b,\ \xi_i} \frac{1}{2}||w||^2 + C\sum_{i=1}^{m}\xi_i \tag{16}$$

$$s.t. \quad y_j(w\cdot x_j + \mathrm{b}) \geq 1-\xi_i, \xi_i \geq 0, i = 1,2,\cdots N \tag{17}$$

The prediction results of the SVM model established in this paper are shown in Figure 6 (taking 2012 data as an example) and Table 4.

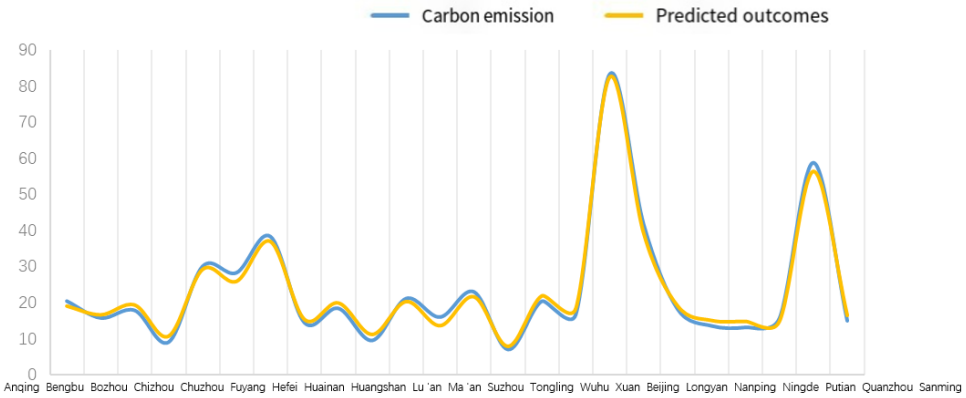

Figure 6: SVM prediction results (2012)

Table 4: Random Forest Regression Results

	MSE	RMSE	MAE	MAPE	R^2
Training set	0.015	0.21708037	0.031	0.09	0.954
Cross verification set	0.023	0.26880593	0.034	0.092	0.911
Test set	0.023	0.27228952	0.036	0.098	0.906

2.7. Model Comparison

The ARIMA time prediction model shows a satisfactory R^2 value of 0.782, which indicates that it is reliable in predicting time series data.The LSTM prediction model performed well, with an R^2 value of 0.864, demonstrating strong predictive power for complex sequence data. In our experiments, the LSTM model stands out in the prediction task with its excellent performance, showing higher accuracy and generalization ability.According to the results, the LSTM prediction model performs best in R^2, reaching 0.864, with higher prediction accuracy. The random forest model performed well on the training set, but decreased slightly on the test set. The SVM regression model obtained a high R^2 value on the cross-validation set. The ARIMA time prediction model ranked third overall.

3. Conclusions

This study discusses the application of carbon emission model, SVM regression analysis, random forest, LSTM neural network model and ARIMA model in the field of carbon emission prediction in environmental science. Among them, the significance of random forest features shows that the policy impact coefficient accounts for 32.5%, indicating that the impact of government policies on carbon emissions is huge, and government policies can play a key role in reducing carbon emissions in the future. The ARIMA model is truncated from the autocorrelation graph at the q order, and the autocorrelation graph is partial to the tail, indicating that the correlation between the data is good, and the time series fitting R^2=0.782, and the fitting effect is good. The LSTM model is also superior in

prediction results, showing strong prediction ability for complex sequence data. Based on the four models, the R^2 of SF model is 0.876, LSTM model is 0.864, ARIMA model is 0.782, and SVM model is 0.954. The SVM model has the smallest error and the best fitting effect.

References

[1] Tan Dan, Huang Xianjin. Correlation analysis and comparison between economic development and carbon emission in eastern, Central and western China [J]. China Population, Resources and Environment, 2008, (03): 54-57.

[2] Wang Feng, Wu Lihua, Yang Chao. Research on driving factors of carbon emission growth in China's economic development [J]. Economic Research Journal, 2010, 45(02): 123-136.

[3] Yang Qian, Liu Huajun. Decomposition and influencing factors of regional differences in carbon dioxide emissions in China: a study based on provincial panel data from 1995 to 2009 [J]. Journal of Quantitative and Technical Economics, 2012, 29(05): 36-49+148.

[4] Yuan Yuan, Xi Qiangmin, Sun Tieshan, et al. The effect of industrial structure on regional carbon emission: An empirical analysis based on multi-country data [J]. Geographical Research, 2016,35(01):82-94.

[5] Chen Zhanming, Wu Shimei, Ma Wenbo, et al. Analysis of influencing factors of carbon dioxide emissions in cities above prefecture level in China: based on the extended STIRPAT model [J]. China Population, Resources and Environment, 2018, 28(10): 45-54.

[6] He Wenju, Zhang Huafeng, Chen Xiongchao, et al. Empirical study on population density, industrial agglomeration and carbon emission in China: Based on agglomeration economy, crowding effect and spatial effect [J]. Nankai Economic Research, 2019, (02): 207-225.

[7] Du Haibo, Wei Wei, Zhang Xueyuan, et al. Spatial-temporal pattern evolution and influencing factors of carbon emissions from energy consumption in the Yellow River Basin: Based on DMSP/OLS and NPP/VIIRS night light data [J]. Geographical Research, 2021, 40(07): 2051-2065.

[8] Deng Rongrong, Zhang Aoxiang. Impact and mechanism of urban digital finance development on carbon emission performance in China [J]. Resources Science, 2021, 43(11): 2316-2330.