ISSN: 2710-0723

Analysis of Factors Affecting Carbon Emissions based on Gray Correlation Analysis

Yixin Yang[1, *], Deyu Li[2]

[1] School of Management Science and Engineering, Anhui University of Finance and Economics, Bengbu, China

[2] School of Economics, Anhui University of Finance and Economics, Bengbu, China

*1438122537@qq.com

Abstract

In order to explore the influencing factors of carbon dioxide emissions, this paper collects data on carbon emissions and several influencing factors that may be related to its changes from 2004 to 2021, considering that too many indicators may make the data redundant and lack of representativeness, using the K-Means clustering algorithm to classify the indicators and defining the names of the corresponding first-level indicators, and analyzing the elbow diagrams for the determination of the model K-value. For the determination of the K value of the model, the optimal number of clusters was analyzed using the elbow diagram, so that the center of each cluster as the first-level indicator data was spliced with the normalized emission data to form a new dataset, and the correlation coefficients between the first-level indicators and the carbon emissions were finally calculated using the grey correlation analysis method. The results show that the ability to mitigate carbon dioxide content has the greatest impact on carbon content, and the correlation coefficient between the two reaches 0.756, which indicates that increasing the green area is very effective in achieving the goal of energy saving and emission reduction.

Keywords

Carbon Emissions; K-Means Clustering; Gray Correlation Analysis.

1. Introduction

In 2022, China issued the Comprehensive Work Program for Energy Conservation and Emission Reduction in the 14th Five-Year Plan. The program approached the issue from various angles, aiming to achieve a number of general goals such as a 13.5% decrease in energy consumption per unit of gross domestic product (GDP) nationwide by 2025 compared to 2020. The release of this program reflects the great importance that the state attaches to energy conservation and emission reduction, and how to achieve the goal has become a matter of concern for many people.

The matter of energy saving and emission reduction can not be delayed, and if we want to better protect the ecological environment, reduce air pollution and promote sustainable economic development, we must understand its nature and find a reasonable solution. Zhang Zhaohui [1] and others use STIRPAT to expand the model to predict carbon emissions and various influencing factors; Lu Haiyue [2] and others synthesize the ecological model, statistics and GIS spatial analysis and other methods to study the spatial and temporal dynamics of carbon change in multiple regions; Jiao Liudan [3] and others mainly use convolutional neural networks to predict carbon emissions from the transportation industry.

ISSN: 2710-0723

In this paper, only from the perspective of emission reduction, in the context of the dual-carbon target, we expect to find ways to effectively reduce carbon emissions, and use gray correlation analysis to analyze the degree of influence of different factors on carbon dioxide emissions from multiple perspectives, with a view to providing a detailed direction for future planning.

2. Data Collection and Processing

2.1. Data Collection

In order to better understand the data related to China's carbon dioxide emissions, the article collected real-time data on emissions from the Carbon Monitor website from January 1, 2019, to September 30, 2023, and drew bar charts to visualize the data as shown in Figure 1.

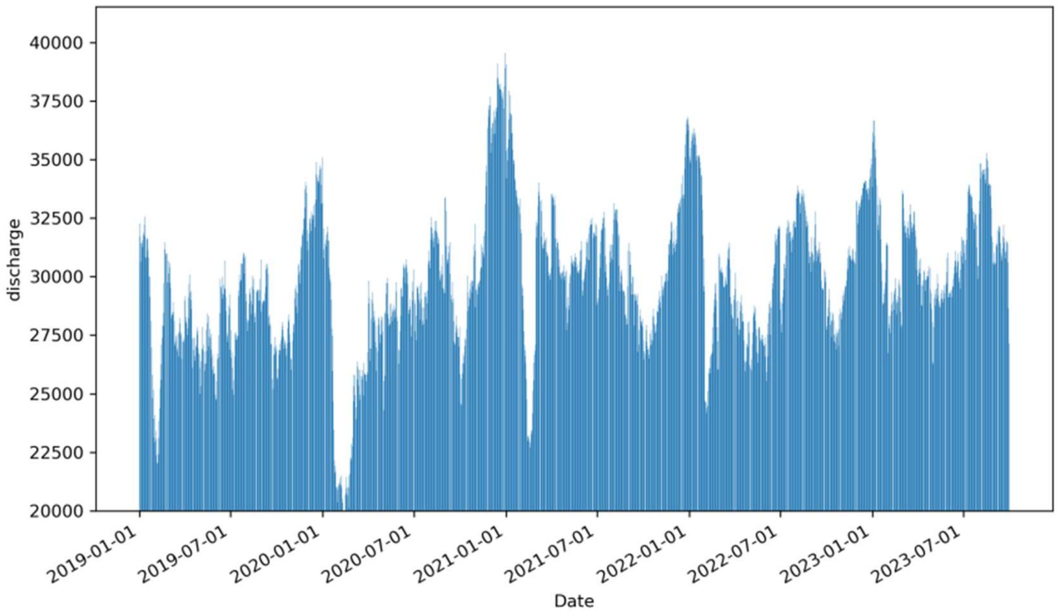

Figure 1. Changes in carbon dioxide emissions

It can be seen that China's daily carbon dioxide emissions are high, between 2019 and 2023, the daily emissions are more than 20,000, the highest is close to 40,000, and reducing carbon emissions is still a huge project.

Then, we obtained the data values of carbon emissions in each sector from CEADs database [4-7] from 2004 to 2021, and collected from the National Bureau of Statistics (NBS) to make a dataset of seven indicators that may affect the content of carbon emissions, namely, the total afforestation area, the forest area, the urban green space area, the total population at the end of the year, the total energy consumption, the private automobile ownership, and the total primary energy production, in the same year.

2.2. Normalization

Because the different index data have different scales, which can easily lead to the dominance of certain features in the model training process, so the data set is normalized to reduce the influence of the scale and accelerate the model convergence speed. Compared with the normalization process, although both ways can be used to solve the problem of the outline, but the normalization will be adjusted to the range of all the data between 0 and 1, which is more suitable for algorithms that need to calculate the distance of the situation, the formula for normalization is as in Equation (1).

ISSN: 2710-0723

$$X = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{1}$$

Where, Xman and Xmin denote the maximum and minimum values of the data in the metric, respectively.

3. Model Construction

Due to the large number of indicators collected, calculating the correlation between each indicator and CO2 emissions one by one lacks representativeness, K-Means clustering method [8] is considered to be utilized to classify the indicators, so that they can be used as secondary indicators to redefine the primary indicators. Between clusters, the elbow diagram should be used to find the optimal number of clusters to facilitate accurate classification.

3.1. Elbow Diagram

The elbow diagram can help us visualize how the clustering cost changes as the number of clusters specified in the K-Means model changes. As the number of clusters increases, the rate of decrease of the cost slows down significantly after a certain point, which is usually regarded as the "elbow point", i.e., the optimal K-value point. The elbow plot using Python code is shown in Figure 2.

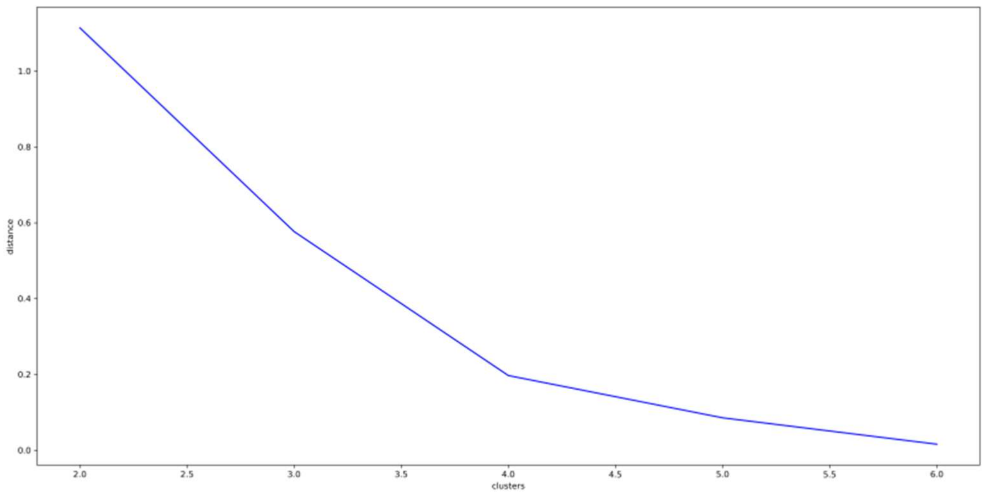

Figure 2. Elbow Method

Based on the information in the figure, it can be found that when the number of clusters is 4, the fold line takes an obvious turn, so set the parameter K in the clustering model as 4.

3.2. K-Means Clustering Model

The core objective of the K-Means clustering algorithm is to minimize the sum of the Euclidean squared distances from each point to the center of the cluster, and it uses formula (2) to find the optimal solution using an iterative approach.

$$J = \sum_{i=1}^{k} \sum_{x \in c_i} \| x - \mu_i \|^2 \tag{2}$$

ISSN: 2710-0723

Where, k is the number of clusters and $\| x-\mu_i \|^2$ denotes the Euclidean square distance from the data point x to the center of mass μ_i.

The specific classification results obtained after the code calculation are shown in Table 1.

Table 1. Classification of indicators

Level 1 indicators	Secondary indicators
Basic Consumptive Capacity	Area of urban green space
	Total population at the end of the year
	Total energy consumption
	Total primary energy production
Mitigation capability	Total afforestation area
CO2 absorption capacity	forest area
Human capacity to influence	Private car ownership

The urban green space area, total population at the end of the year, total energy consumption and total primary energy productivity are combined as the CO2 base consumption capacity, the total area of afforestation is categorized as the capacity to mitigate the CO2 content, the forest area can be regarded as the absorptive capacity, and the private automobile ownership is regarded as the anthropogenic impact capacity.

Then the mean value of all the stores in the cluster is calculated to obtain the cluster center using equation (3), which is used as the first-level indicator data for the purpose of dimensionality reduction.

$$\mu_i = \frac{1}{|C_i|}\sum_{x \in C_i} x \tag{3}$$

Where, |Ci| is the number of data points in cluster Ci.

Table 2. Various types of cluster centers

Years	Cluster 1	Cluster 2	Cluster 3	Cluster 4
2004	0.000081	0.469217	0	0
2005	0.086286	0.419771	0	0.014852
2006	0.123623	0.021485	0	0.034522
2007	0.226718	0.039024	0	0.056528
2008	0.268841	0.407199	0	0.081869
2009	0.334601	0.638267	0.489521	0.125383
2010	0.415345	0.54858	0.489521	0.180664
2011	0.500276	0.570642	0.489521	0.236929
2012	0.562107	0.468635	0.489521	0.29821
2013	0.608072	0.596969	0.489521	0.365622
2014	0.653687	0.456882	1	0.440111
2015	0.690034	1	1	0.511441
2016	0.711492	0.877793	1	0.601879
2017	0.771157	0.999239	1	0.690442
2018	0.83371	0.902215	1	0.773936
2019	0.890362	0.925328	1	0.852332
2020	0.933779	0.809128	1	0.924572
2021	1	0	1	1

ISSN: 2710-0723

The specific values calculated for each type of cluster center are shown in Table 2, which are then merged with the normalized CO2 emissions data to form a new dataset and the correlation is calculated.

4. Calculation of Correlation Degree

In the case of a small sample size, for the analysis of the association between each level of indicators and carbon emissions, it is more appropriate to use gray correlation analysis to calculate. Gray correlation analysis [9] first use formula (4) to calculate the correlation coefficient, compare the proximity of the two series at the same moment, and then use formula (5) to average the correlation coefficient as the correlation between the two indicators.

$$\xi_i(k) = \frac{\min\min|x_0(k)-x_i(k)| + \rho * \max\max|x_0(k)-x_i(k)|}{|x_0(k)-x_i(k)| + \rho * \max\max|x_0(k)-x_i(k)|} \tag{4}$$

Where, x0(k)-xi(k) is the difference between the reference sequence and the comparison sequence at the moment k; ρ is the resolution coefficient, which generally takes values between 0 and 1.

$$\gamma_i = \frac{1}{n}\sum_{k=1}^{n}\xi_i(k) \tag{5}$$

The higher the correlation coefficient, the higher the correlation degree, which indicates the higher similarity of the two indicators, i.e., the greater the influence of the comparative series on the reference series, and it is more indicative that the indicator plays a key role in controlling CO2 emissions.

5. Conclusion

After code calculation, the correlation degree of each level of indicators with carbon emissions data is obtained. The correlation between carbon emissions and itself is regarded as 1, and the heat map is drawn to visualize the results as in Figure 3.

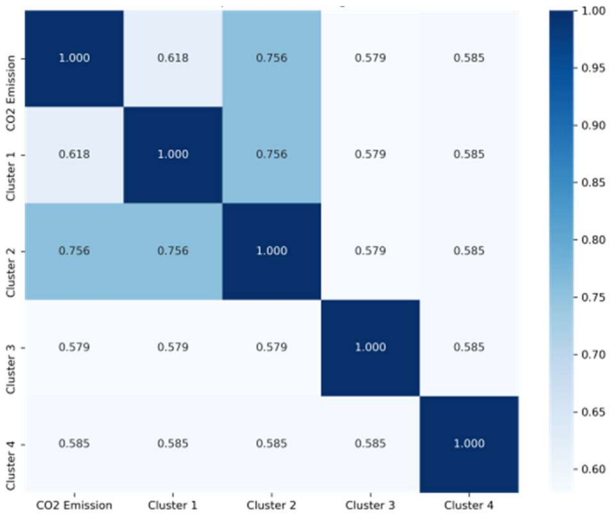

Figure 3. Heatmap of Relational Degrees

ISSN: 2710-0723

From the figure, it can be seen that the correlation between base consumption capacity and carbon dioxide emissions is 0.618, the correlation between carbon dioxide mitigation capacity and the reference series is the highest at 0.756, the influence of carbon dioxide absorption capacity is the lowest, with a correlation of only 0.579, and the correlation coefficient between anthropogenic influences and carbon emissions averages 0.585.

This indicates that the increase of greening area can effectively reduce the carbon dioxide content in the air, which plays a vital role in the energy saving and emission reduction policy, while the use of clean energy instead of oil, coal and other energy sources can also help to slow down carbon dioxide emissions.

Acknowledgments

This work is supported by Anhui University of Finance and Economics National College Student Innovation and Entrepreneurship Training Program Project (Grant No: 202210378298).

References

[1] ZHANG Chaohui, ZHANG Jingya, YU Shiqi. Carbon peak prediction for provinces along the Silk Road Economic Belt from a multi-scenario perspective[J/OL]. Ecological Economy:1-16.

[2] LU Haiyue, QI Jiaojiao, YE Yanlei et al. Characteristics of spatial and temporal changes in carbon balance at different administrative scales in China[J/OL]. Environmental Science:1-29.

[3] Jiao Liudan, Liu Ying, Wu Ya et al. Carbon emission prediction in transportation industry based on convolutional neural network[J/OL]. Railway Transportation and Economy:1-9Shan et al. (2018) "China CO2 emission accounts 1997-2015. Scientific Data", https://www.nature. com/articles/ sdata2017201.

[4] Shan et al. (2020) "China CO2 emission accounts 2016-2017. Scientific Data", https://www.nature. com/articles/s41597-020-0393-y.

[5] Guan et al. (2021) "Assessment to China's recent emission pattern shifts. Earth's Future", https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2021EF00224.

[6] Xu et al. (2024) "China carbon emission accounts 2020-2021. Applied Energy", https:// www. sciencedirect. com/science/article/pii/S0306261924002204#ac0005.

[7] Wang W. Research on smoke detection algorithm combining Kmeans clustering and multispectral thresholding[J]. Industrial Heating,2022,51(04):45-47+57.

[8] LI Shangge,FENG Xueliang. Research on cost risk management of TBM construction based on entropy right and gray correlation[J/OL]. Water Resources Development Research:1-10[2024-02-25].