Research on the correlation between digital economy and carbon emission based on PSO-Kmeans algorithm

Linjing Wu, Yilong Liao, Yi Chen and Yuanbo Liu *

China West Normal University, Sichuan, China

* Corresponding author: liuyuanbo@cwnu.edu.cn

Abstract. Based on the PSO-Kmeans algorithm, this paper deeply explores the correlation between digital economy and carbon emissions in China's prefecture-level cities. Firstly, random forest algorithm is used to deal with missing values, logarithmic transformation is used to improve data skew distribution, and principal component analysis is used to reduce data dimension, which improves the quality of data analysis. Furthermore, K-means clustering algorithm combined with Hopkins statistics was used to verify the feasibility of clustering. Monte Carlo simulation and PSO-Kmeans algorithm were used to optimize the clustering results, and the type characteristics of carbon emission trends in each city were obtained.

Keywords: Random forest algorithm, PSO-Kmeans algorithm, PCA.

1. Introduction

With the continuous progress and popularization of information technology, the digital economy has developed rapidly. In the past course of economic development, the construction of new competitive advantages and new development pattern of the country with digital economy as the core has gained extensive academic consensus [1]. In 2022, the scale of China's digital economy will reach 50.2 trillion yuan, with a year-on-year nominal growth of 10.3%, accounting for 41.5% of GDP 1[2], "The scale of China's digital economy will exceed 55 trillion yuan in 2023, accounting for more than 41.5% of GDP, ranking second in the world [3]. The digital economy has become the core driving force for economic growth in recent years.

In terms of the construction method of digital economy index system, Liu Chuanming (2020) uses the urban digital economy development index data published by the big data platform "Tencent Internet +" to measure the development level of digital economy [4]. Another method is to build a multi-dimensional comprehensive index system for comprehensive evaluation from the perspective of the system. For example, Liu Jun et al. (2020) [5] constructs an evaluation index system for the level of digital economy in China's provinces by constructing three dimensions of digital transaction, Internet and informatization development. For the study on the impact of digital economy on carbon emissions, Jin Dianchen et al. (2023) [6], based on the panel data of 275 prefecture-level cities in China from 2011 to 2021, found an inverted "U-shaped" relationship between the development of digital economy and regional CO2 emissions. Wang Ziyao (2024) [7] focused on the impact and mechanism of digital economy development on carbon emission reduction in 30 provinces in China, while Xu Weicong (2024) [8] proposed the path of digital empowerment for product trade development in countries along the "Belt and Road". Wei Tao et al. [9] Enabling agricultural carbon emission reduction based on digital economy; Deng Rongrong et al. [10] found that digital economy can improve the environment by reducing urban industrial wastewater.

2. Methodology

2.1. Data cleaning and dimensionality reduction

2.1.1. Data cleaning based on random forest

Missing values are common in big data and have a great impact on the data model. According to the missing data, there are also differences in the processing methods. Random forest (RF), as a

machine learning method, has high accuracy and robustness in filling missing values of high-dimensional data. The algorithm flow for filling missing values is as follows:

(1) The column with the smallest proportion of missing values in the data set is extracted, and the missing value data in the remaining columns are interpolated with 0 to obtain a new data set

$$X' = \{x_1', x_2', \dots, x_n'\} \tag{1}$$

(2) Conduct sub-bootstrap sampling on the data set to obtain a training set containing several samples.

(3) Construct a decision tree model $C_i(i = 1,2, \dots, m)$ according to the training set $T_i(i = 1,2, \dots, m)$

(4) Vote on each record according to the classification results

$$C(T) = argmax_C \sum_{i=1}^{m} I(c_i(T) = C) \tag{2}$$

$c_i(T)$ is the first decision tree model, $I(c_i(T) = C)$ is the indicative function of the decision tree, C is the classification label of the prediction of a single decision tree, and $C(T)$ is the final prediction result of all decision trees.

(5) Returns the data set $X = \{x_1, x_2, \dots, x_n\}$, supplementing the missing values in the column.

2.1.2. Dimensionality reduction based on PCA

Due to the correlation between some index data, dimensionality reduction of the original data set is needed to reduce the overfitting problem caused by the large number of eigenvalues, reduce the amount of data computation, and improve the execution efficiency of the algorithm.

Generally, take the first, second..., the KTH principal component. In this paper, the average contribution rate of 90% is taken as the dividing standard.

The greater the contribution rate of the index, the more information the component has integrated, and the more representative it is in its own category.

2.2. Analysis based on K-means clustering algorithm

2.2.1. Hopkins statistics to determine the clustering feasibility

Hopkins statistics are used to evaluate the clustering tendency of a given data set by measuring the probability that it is generated from a uniform data distribution.

For the Hopkins statistic H, when H<0.5, there is unlikely to be a statistically significant cluster in the dataset; When the clustering condition exists in the data set, H will be close to 1. Therefore, we conduct Hopkins statistical test. If the Hopkins statistic value is close to 1, then the data set under study is cluster able data.

2.2.2. K-means clustering algorithm

K-means clustering algorithm is an unsupervised clustering learning algorithm, which divides the data set into different K clusters, and minimizes the sum of distances between each object and its cluster center through calculation. The commonly used distances are: Euclidean distance, Manhattan distance, cosine similarity, Chebyshev distance, etc., and the clustering results are continuously optimized through iteration until the end. The basic principle of the algorithm is as follows:

(1) Define the distance function: For two n-dimensional vectors x and y, the various distances between them are defined as:

Euclidean distance function:

$$d(x, y) = \sqrt{\sum_n^{i=1}(x_i - y_i)^2} \tag{3}$$

Manhattan distance function:

$$d(x, y) = \sum_{i=1}^{n}|x_i - y_i| \tag{4}$$

Cosine similarity function:

$$cos\theta = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \tag{5}$$

Chebyshev distance function:

$$d(x, y) = max(|x_i - y_i|) \tag{6}$$

(3) centroid calculation of the cluster: For a cluster c, its centroid can be obtained by calculating the mean of all samples in the cluster, namely:

$$centroid = \frac{1}{|C|}\sum x(x \in C) \tag{7}$$

(3) Allocation of samples: For a sample x, we need to find its nearest center of mass and assign it to the corresponding cluster:

$$c(x) = \arg \min d(x, centroid_i) \tag{8}$$

2.3. PSO-Kmeans clustering particle swarm optimization algorithm

Pso-kmeans clustering particle swarm optimization algorithm aims to optimize and improve the convergence speed of K-means algorithm through PSO algorithm, make up for the problem that the original algorithm can only achieve local optimization, and improve the global optimization ability of the algorithm. Moreover, K-means clustering algorithm converges faster when seeking the local optimal solution, while PSO algorithm converges less quickly when approaching the optimal solution.

At the initial stage of the algorithm, the initial location of the cluster center required by K-means can be quickly searched by executing the PSO algorithm, which is the optimal initial value of the cluster center. After several iterations by reducing PSO inertia factor, the global extreme value of particle swarm is finally used as the clustering center of K-means algorithm. For the I-th particle, its fitness is defined as the sum of squares of its Euclidean distance from the central point

3. RESULTS

3.1. Data cleaning result

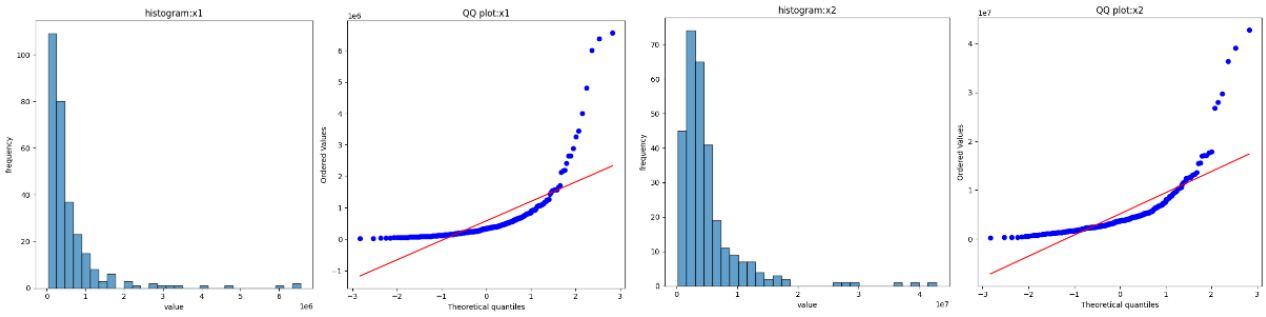

Figure 1. Distribution of X_1 and X_2 before conversion

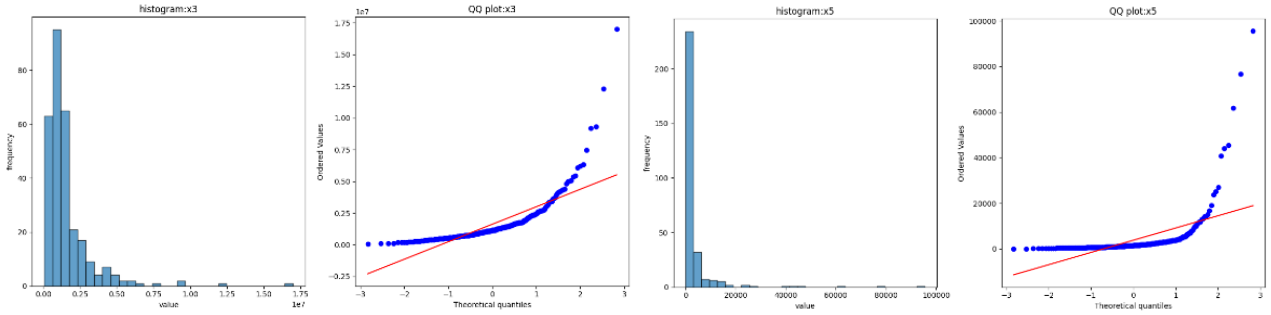

Figure 2. Distribution of X_3 and X_5 before conversion

Since the data of each indicator of digital economy presents a skewed distribution (Figure 1 and Figure 2 shows the skewed distribution of indicators X_1, X_2, X_3 and X_5), it is not suitable for statistical analysis. Practice has proved that some skewness data can be transformed into lognormal distribution by logarithmic transformation of observed values, and its normal value range can be calculated by using the area distribution rule under the normal curve.

3.2. Dimensionality reduction analysis

KMO and Bartlett tests determine the feasibility of applying principal component analysis

KMO and Bartlett tests are the prerequisite tools used to judge whether principal component analysis can be performed. Among them, the detection and evaluation standards of KMO value are shown in the following table:

Table 1. Evaluation criteria of KMO value

KMO	>0.8	0.7~0.8	0.6~0.7	0.5~0.6	<0.5
The situation	very suitable	generally suitable	not very suitable	poor	very unsuitable

For the Bartlett test decision, if the P-value is less than 0.05, the null hypothesis is rejected, indicating that principal component analysis can be used; otherwise, the null hypothesis is not rejected, indicating that the variables are not linearly correlated and can provide information independently, so principal component analysis is not suitable.

Table 2. KMO test and Bartlett test results

KMO		0.943814
Bartlett	Chi-square value	10687.25
	P	0

The KMO test results showed that the KMO value was 0.943814; meanwhile, the Bartlett sphericity test results showed that the significance P value was 0, the horizontal significance was presented, the null hypothesis was rejected, the variables were correlated, and the principal component analysis was effective and the degree was appropriate.

The gravel map is drawn using the data variation of principal components, combined with the characteristic value slope and variance interpretation, which helps us to select the number of indicators within the contribution rate, find the inflection point with the highest cost performance, and determine the number of principal components.

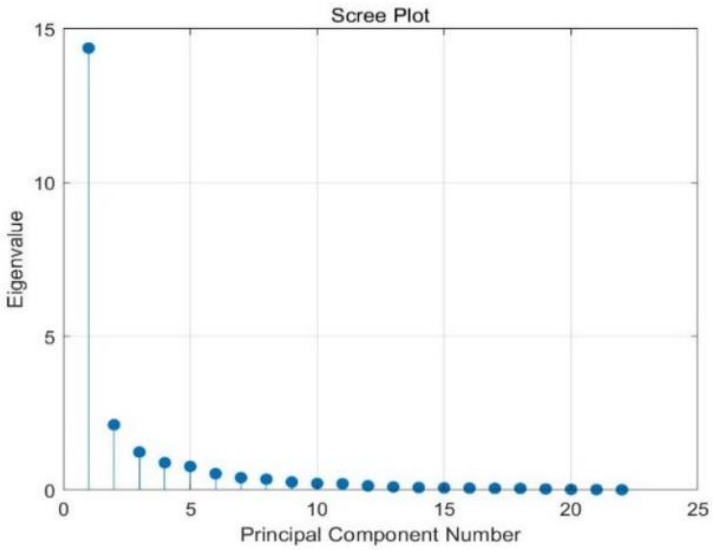

Figure 2. Lithotripsy diagram

To sum up, combined with the contribution rate obtained from the principal component analysis and the results of the above gravel map, this paper finally selected principal components P1, P2, P3, P4, P5, and P6 as the six principal components of the digital economy development index.

3.3. The correlation between urban digital economy and carbon emissions

The correlation results are shown in the figure below:

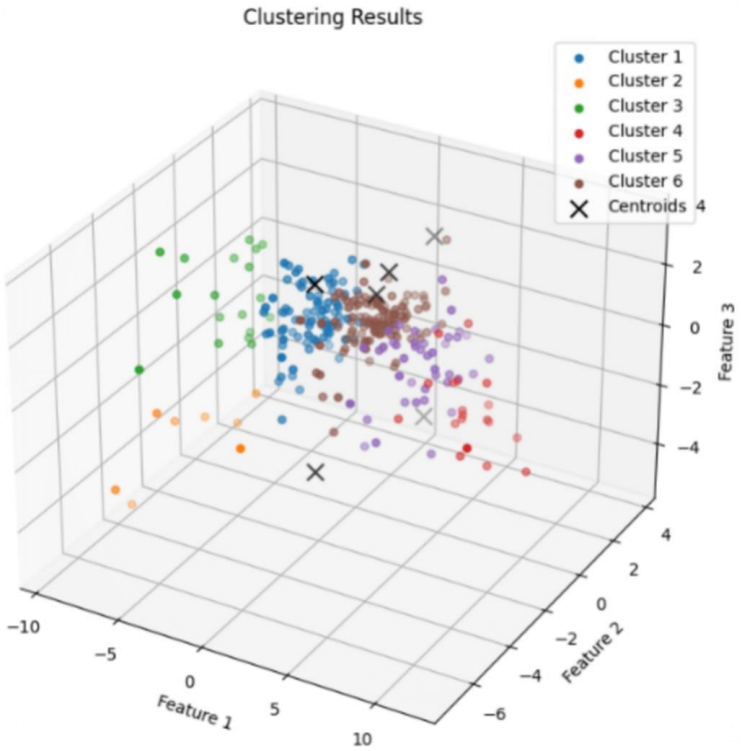

Figure 3. PSO-Kmeans clustering result

The results show that the correlation between urban digital economy and carbon emissions in Chinese cities can be roughly divided into six categories, which can be summarized as high-carbon digital transformation cities, high-carbon traditional economic cities, digital green frontier cities, low-carbon underdevelopment cities, low-carbon population loss cities and low-carbon resource-dependent cities.

4. Conclusion

According to the results, China's prefecture-level cities are finally divided into six categories, which can be summarized as high-carbon digital transformation cities, high-carbon traditional economic cities, digital green frontier cities, low-carbon under-development cities, low-carbon population loss cities and low-carbon resource-dependent cities according to their characteristics.

References

[1] Cui Rong, Zhai Lingyu, Sun Yanan. Network structure and influencing factors of China digital economy spatial correlation [J]. Journal of economics and management review, 2023, 33 (6) 6: 95 - 108.

[2] According to the China Digital Economy Development Research Report (2023).

[3] Zhu Guanghan. (2024). The scale of national digital economy is expected to exceed 55 trillion yuan in 2023. Daily Business Newspaper. (2024 - 01 - 03).

[4] Liu Chuanming, Yin Xiu, Wang Linshan. Regional differences and distribution of the economic development of China's digital dynamic evolution [J]. Science and technology of China BBS, 2020 (03): 97 - 109.

[5] Liu Jun, Yang Yuanjun, Zhang Sanfeng. The digital economy measure of China and the driving factors [J]. Journal of Shanghai economic research, 2020 (6): 81 - 96.

[6] Jin Dianchen, Chen Xin, Liu Shuai. How the digital economy contributes to the "dual carbon" goal: An empirical test based on 275 prefecture-level cities [J]. Guizhou social science, 2023 (9): 134 - 143.

[7] Wang Ziyao. Study on the impact and mechanism of digital economy development on carbon emission reduction [J]. China Price, 2024 (04): 114 - 118. (in Chinese).

[8] Xu Weicong. Mechanism, challenge and path of digitalization enabling high-quality development of agricultural trade between China and countries along the Belt and Road [J/OL]. Contemporary economic management: 1-17 [2024 - 05 - 11].

[9] Wei Tao, Cao Xiaocan. Effects of digital economy development on agricultural carbon emissions in China [J]. Journal of Lanzhou Institute of Technology, 2019, 31 (02): 105 - 110. (in Chinese).

[10] Deng Rongrong, Zhang Aoxiang. China city digital research on mechanism of the effect of environmental pollution and economic development [J]. Journal of southern economy, 2022 (02): 18 to 37.