Regionalization of Transportation Energy Consumption in China Based on Agglomerative Clustering Technology

Xu-chen FANG*, Xu WU and Wei PENG

No. 3 Shangyuan Village, Haidian District, Beijing, China

*Corresponding author

Abstract. Faced with the regional differences in the energy consumption of China's transportation and the unbalanced development trend, the provinces and cities in China are re-clustered based on scientific and rational regionalization techniques to form the clustering results that are most conducive to the formulation of different traffic energy conservation policies in different regions. The research index involves a constructed corridor network, performance of transportation capacity, and energy intensity of passenger and freight transportation in provinces. With average data of 2005–2015 as a sample, 12 indexes and regionalization are analyzed through principal component analysis (PCA) and first-order complete linkage clustering (CLK) methods, respectively. The first-order-CLK method achieves significant results with homogeneity as a measure of the division process and results. With regard to homogeneity and number of provinces in a region, a ten-region-scenario is selected as the final division scenario because this scenario conforms to the algorithm design requirements and is suitable for practical application.

Introduction

The transportation industry is a resource-occupying and energy-consuming industry. With the rapid development of the economy and acceleration of urbanization, the scale of transportation energy consumption increases annually. According to statistics, the proportion of the total energy consumption of the transportation industry accounts for more than 7% of the total annual energy consumption, which has risen significantly in recent years. In 2015, the transportation energy consumption reached 383.176 million tons of standard coal, accounting for 8.91% of the total. Due to the different development status of China's different provinces and cities in terms of economic level, road network construction, and transportation technology level, the huge energy consumption in the transportation industry has forced the authorities to carry out corresponding measures in different regions. However, there is very little research on the division of energy consumption regions. Regional division based on energy consumption is particularly important for the effective implementation of energy conservation and emission reduction policies.

The latest regionalization method series includes the regionalization method that involves dynamically constrained clustering and partitioning proposed by D. Guo. This series covers the combination of three commonly used hierarchical clustering methods (i.e., single linkage clustering, average linkage clustering, and complete linkage clustering) and two contiguity-constraining strategies. This method is effective in studying the regionalization of the presidential election in the United States [1]. R.M. Assuncao et al. proposed an effective regionalization technique that uses the socio-economic geographic units of the minimum spanning tree to effectively reduce the computational cost of AZP to suit the calculation of large datasets [2]. However, in the regionalization of transportation energy consumption, relevant studies have a certain limitation. For example, Yuan from Chang’an University used the system clustering method to study the regional division of transportation in China, but the index selection lacked hierarchy, and the purpose of division was not explicit; therefore, a favorable application effect was not obtained [3]. Wang from Chang’an University used the method of system clustering to study the regionalization of carbon emission reduction in China’s transportation industry, but the region size was uneven, and an unconnected phenomenon was observed in the region [4].
The focus of this study includes the selection of indicators, data processing, and the application of partitioning methods. First, according to the statistical yearbook, the main factors affecting the traffic energy consumption are summarized and analyzed. Based on sufficient indicators, preliminary data processing is carried out to further expand the regional distribution of traffic energy consumption in 30 provinces and cities in China. The division process is based on the first-order-CLK method that involves merging and redistribution. The result ensures high connectivity among the regions and is conducive for analyzing the factors that affect regional energy consumption.

**Indicators Selection and Data Processing**

**Selection of Regionalization Indicators**

Transportation energy consumption is mainly embodied in the operation of the transport sector and includes fuel, coal, gas, and electricity for vehicles. Therefore, this section selected the total energy consumption and transportation energy intensity of passengers and goods as a direct index to reflect the energy consumption of the industry.

With the construction and improvement of the traffic infrastructure, the upper bound of transport capacity and the transport demand caused by the transportation system increase correspondingly and thus indirectly affect the transportation energy consumption. Therefore, this section adopts highway mileage, rail mileage, and density of rail and highway networks as indirect indicators that affect energy consumption. According to the positive externality of transportation, transportation has brought significant economic benefits to the society. Meanwhile, economic development has promoted the transport shifting of passengers and goods. Therefore, this section uses per capita GDP, the volume of passenger and freight transport, and the volume of passenger and freight turnover in each province as an indirect index to reflect the transportation energy consumption.

All these data are obtained from the National Bureau of Statistics, China Transportation Yearbook, and China Energy Statistical Yearbook. The turnover volume of passenger and freight is calculated with the passenger and freight conversion coefficient stipulated in the China statistics system. The coefficient of railway, coastal, and inland river transportation is 1, that of highway transport is 0.1, and that of aviation is 0.073.

This study did not analyze the cases of Tibet, Hongkong, Macao, and Taiwan because some index data for particular years in the Tibet Autonomous Region are missing, and the National Bureau of Statistics does not consider the data of Hong Kong, Macao, and Taiwan. Moreover, due to the lack of pipeline transport data in several areas, this study discussed only four modes of transportation, namely, highway, railway, waterway, and civil aviation transport.

**Data Processing**

To reduce the number of indexes while maintaining data integrity, this paper adopts PCA to process the data.

The new principal component is obtained through PCA, and the results are shown in Table 1. Among the 12 principal components, the first four with an eigenvalue of more than 1 are selected to form a new index system. The accumulated variance of the four new indexes accounts for 86.243% of the variance of all principal components. This result is in line with the requirements because the value is greater than 85%. The positive value refers to positive correlation, and the negative value represents negative correlation.

After obtaining the number and ingredient matrix of the new index, the next step is to calculate the new indicator value containing 30 administrative regions. The values of the four new indicators can be calculated as follows:

\[
Z_{ki} = \sum_{j=1}^{12} F_{kj} X_{ij} \quad (1)
\]
where $Z_{ki}$ denotes the new index value obtained from the $i$th region of the $k$th new index by the joint action of the 12 original index values; $k=1, 2, 3, 4$ represents four new indicators; $i=1, 2, \ldots, 30$ represents the 30 study areas; $X_{ij}$ represents the normalized value of the $j$th original index value in the $i$th region; and $F_{ij}$ represents the eigenvector, which is equal to the $j$th component load of the $k$th new index in the component matrix divided by the square root of the $k$th initial eigenvalue. With this formula, four new index values can be obtained and used for regionalization.

| Component | Initial Eigenvalue | Extracted Sums of Squared Loading |
|-----------|-------------------|----------------------------------|
|           | Total Variance [%]| Accumulation [%]                 | Total Variance [%]| Accumulation [%] |
| 1         | 4.237             | 35.307                           | 4.237             | 35.307             |
| 2         | 3.661             | 30.51                            | 3.661             | 30.51              |
| 3         | 1.372             | 11.436                           | 1.372             | 11.436             |
| 4         | 1.079             | 8.989                            | 1.079             | 8.989              |
| 5         | 0.604             | 5.033                            | —                 | —                  |
| 6         | 0.447             | 3.729                            | —                 | —                  |
| 7         | 0.236             | 1.967                            | —                 | —                  |
| 8         | 0.129             | 1.074                            | —                 | —                  |
| 9         | 0.089             | 0.742                            | —                 | —                  |
| 10        | 0.074             | 0.62                             | —                 | —                  |
| 11        | 0.044             | 0.366                            | —                 | —                  |
| 12        | 0.027             | 0.228                            | —                 | —                  |

**Regionalization with the First-Order-CLK Method**

The first-order-CLK method used in this work belongs to the agglomerative clustering approach and can generate a spatial adjacency tree. The CLK method defines the distance between two regions as the distance of the data pair with the largest attribute difference and connects the two data points to generate the connection of two regions, which is the branch of the tree. When this method is used in clustering, the algorithm begins searching from a single province that contains data information, finds the two provinces with the largest attribute difference, and connects them to a new provincial cluster. When two clusters contain multiple provincial data points, the distance between them is still defined as the provincial connection with the greatest attribute difference between two clusters. However, the tree does not contain the internally generated provincial branches of the cluster because the inner branches of the cluster produce loops. The distance definition of the CLK can be described as follows:

$$d_{CLK} (A, B) = \max_{a \in A, b \in B} (d_{ab})$$

(2)

where $A$ and $B$ are the two clusters formed during the clustering process, $a$ is the arbitrary data point in $A$, $b$ is the arbitrary data point in $B$, and $d_{ab}$ is the distance between two arbitrary data points.

On the basis of the CLK agglomerative clustering method, the first-order constraint strategy is applied to the clustering process to form the first-order-CLK method. In actual situations, adjacent relations exist among provinces. For example, Fujian (A) is only adjacent to Guangdong (B), Jiangxi (C), and Zhejiang (D). Then, the connection between Fujian (A) and Guangdong (B) is called the first-order edge, and the connection between Fujian (A) and non-neighboring Anhui (E) is the non-adjacent edge. The first-order constraint strategy only considers the first-order edge in the clustering process. That is, only the distance between neighboring provinces is involved in the comparison of the distance between clusters. Therefore, the tree established by the first-order...
constraint strategy is a spatial adjacency tree based on actual adjacency.

After forming the adjacent tree of the province, the tree should be divided scientifically into a high-quality provincial cluster by means of the division algorithm so that the regionalization problem is transformed into a tree partitioning problem. Each time an edge of the tree is deleted in the algorithm, the result contains one more subtree. If \( k \) clusters of provinces need to be generated, then the \( k - 1 \) edge must be removed [5]. The purpose of the division is to reduce the heterogeneity within the region and maintain a large difference between regions. Heterogeneity within the region is determined by the difference in the attribute values of all provinces in the region.

In this work, the sum of the squares of deviations (SSD) is used to describe heterogeneity within the provincial cluster. The more concentrated the attribute value of the province is within the region, the smaller the value of SSD is. SSD can be described as follows:

\[
SSD = \sum_{i=1}^{m} \sum_{j=1}^{n} (x_{ij} - \bar{x}_j)
\]

where \( x_{ij} \) represents the \( j \)th attribute value of the \( i \)th object in the cluster, \( \bar{x}_j \) represents the average of the \( j \)th attribute values of all the provinces in the cluster, \( m \) represents the number of provinces within the cluster, and \( n \) represents the number of attributes contained in each province.

The partitioning algorithm must iterate the segmentation of the tree and then select the best subtree scheme [6]. In this work, the homogeneity of the partition results is defined to represent the difference between the heterogeneity of the originally adjacent tree of the province and the sum of the heterogeneity of all subtrees. Homogeneity can be described as follows:

\[
W = \max\{SSD_{T_0} - SSD_{T_1} - SSD_{T_2} - \cdots - SSD_{T_k}\}
\]

where \( W \) denotes the homogeneity of the final partition scenario, \( SSD_{T_0} \) denotes the heterogeneity in the originally adjacent tree of the province, and \( SSD_{T_k} \) denotes the heterogeneity in the \( k \)th subtree.

**Clustering Result**

In the process of exploring the partition scheme, the homogeneity of the clustering is optimized with the increments in the number of districts. However, in the 11th or more division scenario, some regions contain only one province. In view of the homogeneity and number of provinces in the region, this study selects the ten-region-scenario in all results of first-order-CLK as the final division scenario. The regional coloring is shown in Figure 1.

Under the limitation of the first-order constraint strategy in the clustering process, Figure 1 shows that the final classification scheme consists of 10 regions. Each region in the geographical space is composed of adjacent provinces, thus avoiding the emergence of a region that contains only one province. One of the purposes of this work is to use this method to implement regionalization. This regionalization method is beneficial to the joint management of the current situation of high transportation energy consumption between provinces and to the rationalization of the implementation of transportation policies and resource distribution. The method also increases the practicability of the division results.

**Conclusion**

(1) The spatial distribution of transportation energy consumption has its own intrinsic features, which ultimately determine the division of energy consumption regions.
(2) The regional division of traffic energy consumption is the basis for all regions in the country to formulate energy conservation and emission reduction policies, and also serves as the basis for the future regional joint transportation governance.

(3) Based on regional homogeneity and inter-regional heterogeneity, the country is divided into ten energy-consuming regions using scientific division methods.

![Figure 1. Ten-region-scenario derived from the first-order-CLK method.](image)

**Acknowledgments**

This research was supported by the National Natural Science Foundation of China (No. 61503022), the Fundamental Research Funds for the Central Universities (No. 2015JBM061).

**References**

[1] Guo, D. Regionalization with Dynamically Constrained Agglomerative Clustering and Partitioning (REDCAP). International Journal of Geographical Information Science, Vol. 22, No. 7, 2008, pp. 801-823.

[2] Assuncao, R.M, and Neves, M. C. Efficient Regionalization Techniques for Socio-Economic Geographical Units Using Minimum Spanning Trees. International Journal of Geographical Information Science, Vol. 20, No. 7, 2006, pp. 797-811.

[3] Changwei Yuan, Jianwei Wang, and Wenxin Lin. Regional Division Method of Transportation in China. Journal of Chang'an University, Vol. 26, No. 4, 2006, pp. 79-82.

[4] Jianwei Wang, Ping Li, and Jie Gao. Regional Division of Carbon Emission Reduction in China's Transportation. Journal of Chang'an University, Vol. 32, No. 1, 2012, pp. 72-79.

[5] Ang, B W. and F.Q. Zhang. A Survey of Index Decomposition Analysis in Energy and Environmental Studies. Energy, Vol. 25, No. 12, 2000, pp. 1149-1176.

[6] Fu, Z. and J. Hao. A Three-Phase Search Approach for the Quadratic Minimum Spanning Tree Problem. Engineering Applications of Artificial Intelligence, Vol. 46, No. PA, 2015, pp. 113-130.