Research on the Definition of Small and Medium-sized Coastal Dry Bulk Shipping Enterprises Based on Data Mining

Hongbin Wei¹², Yunbo Li¹*, Jiaye Gong¹ and Wenquan Du¹
¹Shanghai Maritime University, China
²China Waterborne Transport Research Institute, China

*Corresponding author email: liyb@shmtu.edu.cn

Abstract. Shipping industry is an important part of the transportation industry. Small and medium-sized shipping enterprises play an important part in protecting the internal vitality of the shipping industry. However, there is no research on the definition of large and medium-sized shipping enterprises. Data mining method has become a convenient tool for mining the value of large data. This study first defines of shipping enterprises, and then discusses the definition of small and medium-sized coastal dry bulk shipping enterprises for the first time from the perspective of large data. This study is the first time to apply the cluster analysis method to research the data value of shipping enterprises, aiming to quantitatively define the scale of coastal dry bulk shipping enterprises from the perspective of large data. This research is the first time to establish a data mining model based on large data of shipping enterprises. The shipping capacity data, ship data and financial operation data are selected as data variables. With the selected data variables, data standardization, outlier processing and clustering analysis are carried out step by step. Taking the coastal dry bulk shipping enterprises as an example, the model is analysed. Through data mining, we get the main influence data variables that distinguish large and medium-sized coastal dry bulk shipping enterprises, and get the quantitative definition results of coastal dry bulk shipping enterprises. The research results serve as a basis for further analysis of dry bulk shipping enterprises, and the research method can be used as a tool for the scale classification of other shipping enterprises.

1. Definition of Shipping Enterprise

Some researchers in the literature [1] [2] [3] [4] proposed that the positioning of small and medium-sized shipping enterprises is that the number of ships is less than 100 and the total capacity is less than 1 million tons. Some scholars also put forward that the enterprises with total assets of less than 500 million yuan and owned and operated ships of less than 100000 tons are small and medium-sized shipping enterprises. There is no systematic scientific basis for the above definition, not to mention the definition of in-depth subdivision of shipping enterprises. If we study the characteristics, strategies, laws and evaluation of small and medium-sized shipping enterprises, and make management policies, we need to point out the definition of small and medium-sized shipping enterprises. Some researchers in literature [5] [6] [7] [8] [9] proposed that the scale of an enterprise should be measured according to the number of employees in the enterprise, and the scale of the enterprise should be determined according to this number. In this study, a set of larger data mining model method is established, including four steps: data variable selection, data standardization, outlier processing and cluster analysis. Through the verification of the data of the specific shipping enterprises, the quantitative numerical basis for the division of large and medium-sized and small-sized coastal shipping areas and dry bulk shipping enterprises is obtained. It is used as a basis for further study of the characteristics of small and medium-sized shipping enterprises.
sized coastal dry bulk shipping enterprises. At the same time, the model method can also be used to study the definition of other shipping enterprises.

2. Data Mining Method

In the aspect of large data mining, clustering analysis method has been widely used. See [13] [14] for details.

2.1. Select Data

Firstly, data of coastal dry bulk shipping enterprises to be mined is selected. Data selection should be closely related to the mining target; at the same time, it can reflect the characteristics of the mining object to a certain extent; the number of dimensions of data variables is enough to show significant differences between different research objects; the correlation between data variables is weak. According to the above principles, capacity data, ship data and financial operation data of coastal dry bulk shipping enterprises are selected as data variables. For large data collected and sorted out, the data variables are classified according to the total tonnage of ships, the number of ships and the business income of the enterprise. The correlation coefficient is calculated by the product difference method. Based on the difference between the two variables and their average values, the correlation degree between the two variables is reflected by the multiplication of the two differences. Definition formula of correlation coefficient:

\[ \rho_{xy} = \frac{\text{Cov}(X, Y)}{(D(X)D(Y))^{1/2}} \]  

Where, \( \text{Cov}(X, Y) \) is the covariance of \( X \) and \( Y \), \( D(X) \) is the variance of \( X \), and \( D(Y) \) is the variance of \( Y \). Through the correlation coefficient check, the correlation among the three data variables was very weak. Therefore, data variables were retained as input data for data mining classification. On the basis of big data analysis, after preliminary data screening, 1102 shipping enterprise values of effective data are obtained, including the number of ships corresponding to each enterprise \( YSSS \), total tons \( YSZD \) and financial index data \( YSCW \). See table1.

| Case processing summary. |
|-------------------------|
| N            | Percentage | YSSS | YSZD | YSCW |
| Effective    | 1102       | 100  | 1102 | 1102 | 1102 |
| Missing      | 0          | 0    | 0    | 0    | 0    |
| Total        | 1102       | 100  | 1102 | 1102 | 1102 |

2.2. Data Standardization

Data standardization is preliminary data base processing work of data mining. The data variables of coastal dry bulk shipping enterprises have different physical unit dimensions. The total tonnage of shipping enterprises is tons, but the number of ships of shipping enterprises has no unit, and the tonnage of ships is several orders of magnitude different from the number of ships. In addition to the financial data, the level of the three data variables is very different. If the original data value is directly used for analysis, the role of data variables with higher value in comprehensive analysis will be highlighted, and the role of data variables with lower value level will be relatively weakened. Therefore, in order to ensure the reliability of the results, it is necessary to standardize the original data variables. Logarithm conversion: take the natural logarithm of the data, and use the log() function to realize it. Generally speaking, natural logarithm transformation can make the data range in the range of 0 ~ 1 larger and the data range in the range of > 1 more compact. Z-score standardization, also known as standard deviation standardization, which is based on the mean and standard deviation of the original data. This method is suitable for the case where the maximum and minimum values of sample data are unknown or there are outliers beyond the range of values. The steps of Z-score standardization are as follows: The first step is to find out the arithmetic mean and standard deviation of each data variable; The second step is to standardize:
Among them: $Z_{ij}$ is the standardized variable value, $X_{ij}$ is the actual variable value. After standardization, the variable value fluctuates around 0. If it is greater than 0, it means higher than the average, if it is less than 0, it means lower than the average. After data standardization, the three data variables are transformed into dimensionless pure values, and the order of magnitude is close, which is convenient for subsequent data mining analysis. For the YSSS index data, the data in the range of more than 1 is logarithmically converted to the YSSSB in the range of 0 ~ 1 after standardization; For the YSZD index data, the data in the range of more than 1 is logarithmically converted to the YSZDB in the range of 0 ~ 1 after standardization; For the analysis of YSCW index data, different from YSSS index data and YSZD index data, the data have outliers beyond the value range. The standard deviation standardization method is selected, and the YSCWB after the standardization of YSCW index data is obtained through Z-score standardization; Name of shipping enterprise is replaced by MC serial number. After data standardization, the indicators are easy to handle the subsequent outliers. See table 2.

Table 2. Data standardization.

| Original value | Number of ships | Capacity | Finance |
|----------------|----------------|----------|---------|
| Enterprise     | YSSS           | YSZD     | YSCW    |
| Standardization value | YSSSB | YSZDB   | YSCWB   |

2.3. Outliers Processing
Outlier processing is an important step before data mining. There are data for distortion in the process of data recording, integration, summary and so on, which will produce abnormal values in the sample data. The existence of outliers causes deviation to the accuracy of data analysis results and affects the accuracy of data results. In order to objectively and truly reflect the characteristics of the data, the standardized data are processed with outliers. The judgment of outliers comes from Chebyshev inequality. For any given $\varepsilon > 0$, it has been found:

$$P(|X - E(X)|) \leq \frac{D(X)}{\varepsilon^2}$$

If the population is a general population, the dispersion degree of statistical data and an average value can be reflected by its standard deviation. Therefore, in general, at least 75% of all data are within 2 standard deviations of the average. Of all the data, at least 88.9% are in the range of 3 standard deviations of the average. Considering the particularity of the survey data, this study adopts to eliminate the discontinuous extreme outliers to ensure the accuracy of the results. See figures 1 to figure 6.
After the processing of outliers, the input data for clustering analysis are obtained. 1098 shipping enterprises correspond to the number of ships YSSS, total tons YSZD and financial index data YSCW of each enterprise at the same time. See table 3.

| N   | Percentage | YSSS | YSZD | YSCW |
|-----|------------|------|------|------|
| Effective | 1102      | 100  | 1102 | 1102 |
| Missing   | 0          | 0    | 0    | 0    |
| Total     | 1102       | 100  | 1102 | 1102 |

2.4. Cluster Analysis[15]

Clustering results, put the shipping enterprises with the same or similar properties in the same set, and put the shipping enterprises with different properties in different sets. Euclidean distance measure is utilized in cluster analysis. Euclidean metric (also known as Euclidean distance) is a commonly used definition of distance, which refers to the real distance between two points in m-dimensional space, or the natural length of the vector (that is, the distance from the point to the origin). Euclidean distance in two-dimensional and three-dimensional space is the actual distance between two points. In the study, the distance between sample points is measured by the square Euclidean distance

\[ d_{ij} = \sum_{k=1}^{m} (x_{ik} - x_{jk})^2 \]  

(4)

Where \( d_{ij} \) represents the distance between sample \( i \) and sample \( j \); \( x_{ik} \) Represents the value of sample \( i \) on variable \( k \). The cluster analysis method classifies shipping enterprises according to their individual characteristics. Similar or similar shipping enterprises gather in one class, and different shipping enterprises gather in different classes. The result of classification is that the shipping enterprises in the same size category have great similarity, and the shipping enterprises in different categories have a great
According to various characteristics of shipping enterprises, according to the degree of intimacy in nature, automatic classification is carried out without prior knowledge, and the classification results of large and small shipping enterprises are produced. Through clustering analysis, the classification results of large shipping enterprises and small and medium-sized shipping enterprises are obtained on the basis of no subjective limit.

From Figure 7 and figure 8, we can see the results of cluster analysis.

3. Definition of Small and Medium Coastal Dry Bulk Shipping Enterprises
Select the domestic coastal dry bulk shipping enterprises, take the total tonnage of the coastal dry bulk shipping enterprises as the data variable of transportation capacity, take the number of the coastal dry bulk shipping enterprises as the data variable of ships, take the annual business income of the coastal dry bulk shipping enterprises as the financial data variable, according to the logic of large data mining, respectively carry out large-scale and large-scale operations for the coastal dry bulk shipping enterprises According to the classification of small and medium-sized enterprises, the results are as follows: The annual operating revenue of coastal dry bulk shipping enterprises is the main influence quantity variable of large and medium-sized and small-sized enterprises classification, and the influence of capacity data variable and ship data variable on distinguishing large and medium-sized coastal dry bulk shipping enterprises is weaker than financial operation data variable; The main difference between large-scale coastal dry bulk shipping enterprises and small and medium-sized coastal dry bulk shipping enterprises is the financial operation data variable; The middle-aged business income of the coastal dry bulk shipping enterprises is less than 1.1 billion yuan; Compared with the previous experience, it is believed that the main difference between the large-scale coastal dry bulk shipping enterprises and the small and medium sized coastal dry bulk shipping enterprises is the financial operation data variable. The large-scale coastal dry bulk shipping enterprises have different ship capacity and ship number data It will be larger than the data of small and medium-sized coastal dry bulk shipping enterprises, so it is impossible to judge the size of coastal dry bulk shipping enterprises from the total tonnage or number of ships.

4. Discussion
In the past, small and medium-sized coastal dry bulk shipping enterprises had a qualitative understanding, that is, independent accounting of shipping enterprises, regardless of the size of assets, operating goods, transportation scope, the possession and allocation of resources in the industry are not dominant enterprises called small and medium-sized shipping enterprises. In this study, large and small coastal dry bulk shipping enterprises are classified from the perspective of large data, and the basis of dividing large and small coastal dry bulk shipping enterprises is obtained by using data mining method. Different from the understanding that a large number of ships and a large tonnage of ships are large shipping enterprises, on the basis of comprehensive consideration of the operating income of coastal dry bulk shipping enterprises, the "large" and "small" of coastal dry bulk shipping enterprises are...
scientifically evaluated by large data mining, which avoids the limitations of using a single factor evaluation.

5. Conclusion
The purpose of this study is to discuss the definition of small and medium-sized coastal dry bulk shipping enterprises. In the research process, the large data mining method is used, and a set of data processing and mining method is established. Considering the multi-dimensional data variables such as the scale of transportation capacity and business finance of the coastal dry bulk shipping enterprises, the paper defines the small and medium sized coastal dry bulk shipping enterprises from the perspective of data mining method, and obtains the small and medium sized coastal dry bulk shipping enterprises with an annual operating revenue of less than 1.1 billion yuan.

References
[1] Qian Fang 2005 Discussion on Brand Management of China′s Small and Medium-Sized Shipping Enterprises (Shanghai: Shanghai Maritime University) p 12
[2] Li Shu Liang 2005 Study on The Development Strategy of Small and Medium-Sized Shipping Enterprises in China (Qingdao: Ocean University of China) p 8
[3] Li Yiran 2017 Research on Ship Financing Mode of Small and Medium-Sized Shipping Enterprises (Tianjin: Tianjin University of Technology) p 18
[4] Zhang can 2019 Research on The Development Strategy of Small and Medium-Sized Shipping Enterprises in Northern Anhui Based on SWOT Analysis (Anhui: Anhui University of Finance and Economics) p 12
[5] Mckeown 2017 From entrepreneurship to small-to-medium enterprises: Exploring the challenge and the opportunities Journal of Management & Organization (JOM vol 2) ed Remi Ayoko (London: Cambridge University Press) pp 165-166.
[6] Hollenstein 2005 Determinants of International Activities: Are SMEs Different? Small Business Economics (Small Bus Econ vol 24) ed Zoltan J. Acs (NETHERLANDS: SPRINGER) pp 431-450.
[7] Kumar 1999 What Determines Firm Size? Social Science Electronic Publishing (SSRN vol 1).
[8] Xiu-yin 2012 International Conference on Management Science and Engineering - Annual Conference Proceedings (Dallas) vol 1 (US & CANADA: IEEE) pp 796
[9] Qian Wu and Yinggui Qiu 2019 International Conference on Economic Management and Model Engineering (Malacca vol 1) ed IEEE (US & CANADA: IEEE) p 360
[10] Hastie, T 2009 The Elements of Statistical Learning Data Mining, Inference, and Prediction, Second Edition (New York: Springer E-books)
[11] Sun Yun Woo 2006 Statistics of: Data Collection, Collation and Analysis (Shanghai: Shanghai University of Finance and Economics Press)
[12] Wu Xizhi 2006 statistics: from data to conclusion (2) (Beijing: China Statistical Publishing)
[13] Yang Hao 2013 Application of cluster analysis based on SPSS in industry statistics (Jilin: Jilin University) p 23
[14] Alexander Strehl and Joydeep Ghosh 2002 Cluster ensembles: a knowledge reuse framework for combining partitionings Eighteenth National Conference on Artificial Intelligence ed Rina Dechter ( CA United States: American Association for Artificial Intelligence) pp 93–98
[15] Kaufman L 2009 Finding Groups in Data: An Introduction to Cluster Analysis (Rhineland: DBLP)