National Urban Climate Clustering Analysis for Cluster Analysis of Container Handling Stations

Ling-xi ZHU*, Han BAO and Shi-lin LI

School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, China
*Corresponding author

Keywords: Rail transport, Railway container, Container corrosion, Urban climate clustering, Fusion algorithm.

Abstract. The causes of damage during the operation of the container mainly include the different types of goods, the loading and unloading operations of loading and unloading stations, and the rust damage caused by transportation and staying in different areas. Classification of railway container handling stations according to the climate data of the city in which it is located is extremely important for studying the impact of climate on container corrosion. According to the main factors of container corrosion, this paper selected the corresponding climate indicators, wrote the crawling program, and collected the original climate data of 320 prefecture-level cities in China from the weather website. Through the preprocessing of data, the paper obtained the basic indicator data of urban climate clustering, and completed analysis by Kmeans clustering method. Sun, rain, salt, and pollution become the main factors in the corrosion of containers. And after cluster analysis, 320 cities are divided into 10-15 clustering groups.

Introduction

In recent years, with the continuous opening of the new railway line, the input of railway containers has increased. As of 2015, the number of railway containers held reached 280,000 TEU, and the proportion of railway container transportation to the total volume of railway freight transportation was 2% from 2010. It has grown to 3% in 2015 [1].

Through the analysis of the container utilization data of China Railway from 2007 to 2017, it can be found that the utilization rate of China's railway containers is relatively low, as shown in Figure 1.

![Figure 1. Average number of times of use of railway containers in China over the years.](image-url)
It can be found from the above figure that the frequency of use of railway containers in China is relatively low, especially before 2010, the average number of uses per year is only about four. After 2010, there has been improvement, but the average is only about 11 times. According to the general railway transportation, the average time of each transportation is 10 days, then on average, the transportation time of the container on the way is only about 120 days. This shows that for Chinese railway containers, basically two-thirds of the time is waiting for transportation. In China's container yards, there are a small number of shields. Most of the fire farms are open-air freight yards. In this way, containers are parked at different locations, and the climatic factors at the location will have a significant impact on the corrosion of containers.

However, for areas that are prone to rust, as well as climate conditions at different locations, there is currently a lack of quantitative analysis based on the subjective experience of managers. And because of the different time periods, each region will have a large climate change, which is difficult to judge accurately based on experience. Therefore, the quantitative analysis method is used to obtain the high-incidence area and time period of the container corrosion, which plays an important role in analyzing the corrosion effects of different locations on the container, thereby reducing maintenance costs and improving transportation efficiency.

For the corrosion damage of containers, research has been conducted at home and abroad. Tang Rongjian et al. [2] analyzed the difference in corrosion rate of different parts of the container under the same environment and analyzed the reasons; Liao Guodong et al. [3] exposed the steel plate to the natural environment in the outdoor, shed and warehouse of the three test stations in Xisha, Hainan and Guangzhou. The corrosion test shows the corrosion rate of the three places and points out that the sun, rain, salt and pollution are the main causes of corrosion; Song Fengliang et al. [4] qualitatively analyze the causes of container corrosion and collision damage and propose countermeasures. According to many literatures on steel corrosion, temperature, water, pH and impurities are the main causes of steel corrosion.

Based on these main influencing factors, this paper proposes an index system that uses weather information to cluster cities. The reptile program is used to obtain the raw data of each region of the weather system and process it, and the clustering index values of each region are obtained, and cluster analysis is performed. The classification of railway container handling stations based on climate data is completed by geographically matching 1,200 railway container handling stations and 320 prefecture-level cities.

**Regional Cluster Analysis**

When selecting regional clustering indicators, according to the research status, sun exposure, rain, salt and pollution are the main factors of container corrosion. According to these factors, the principles for constructing cluster analysis indicators are as follows:

- The climate gaps between provinces and cities in China are large. To ensure accuracy, the prefecture-level cities are selected as clustering objects.
- The climate varies greatly from region to region at different times and should be analyzed by time period.
- Sun exposure is reflected in temperature and sunshine hours. Rainfall is reflected in rainfall and rainfall days.
- The main quantitative indicator of pollution is sand dust. According to the available data, the winds and sands of other provinces except Xinjiang are small, and it is difficult to quantitatively count the dust indexes of various prefecture-level cities. Therefore, this article does not conduct research on Xinjiang for the time being, and does not consider pollution when researching other regions.
- pH refers to the pH value of rainfall, which is related to the industry and geographical location of the area. According to the available research data, the pH value of rainfall in most areas is unknown, so it will not be considered.
The degree of influence of rainfall and snowfall on containers is different. Snowfall is difficult to eliminate, and damage to containers is greater than rainfall, so it needs to be considered separately.

For the sun and rain factors, it is also necessary to consider the average number of days. For example, the rate of rusting of steel with continuous rainfall is significantly greater than the rate of rusting between rainfall and sunny days, so the average rainfall duration and average snowfall duration are selected.

In summary, for regional clustering, the spatial granularity used is the prefecture-level city, and the time granularity is the month, because the minimum statistical time period of rainfall is the month, indicators we selected are temperature, number of sunny days, average sunny time, rainfall, number of rain days, average rainfall duration, snowfall days, average snowfall duration, and peak sunshine duration.

Index acquisition path and calculation method. Use the weather network [5] information to obtain weather data, taking Beijing as an example.

| Date       | Maximum temperature [°C] | Minimum temperature [°C] | The weather              | Wind direction | Wind power |
|------------|--------------------------|---------------------------|--------------------------|----------------|------------|
| 2016-06-01 | 30                       | 18                        | cloudy                   | Southwest      | Level 3-4   |
| 2016-06-02 | 31                       | 20                        | sunny to cloudy          | Southwest      | Level 3-4   |
| 2016-06-03 | 29                       | 19                        | cloudy                   | Southwest      | Level 3-4   |
| 2016-06-04 | 30                       | 19                        | cloudy                   | Southeast      | Level 3-4   |
| 2016-06-05 | 30                       | 18                        | cloudy                   | Southeast      | Level 3-4   |
| 2016-06-06 | 31                       | 20                        | cloudy thunderstorm      | Northeast      | Level 3-4   |
| 2016-06-07 | 26                       | 16                        | cloudy to sunny          | Southwest      | Level 3-4   |
| 2016-06-08 | 31                       | 19                        | cloudy                   | Southeast      | Level 3-4   |
| 2016-06-09 | 32                       | 22                        | cloudy thunderstorm      | Southwest      | Level 3-4   |
| 2016-06-10 | 31                       | 20                        | cloudy to sunny          | Northeast      | Level 3-4   |
| 2016-06-11 | 30                       | 18                        | cloudy to sunny          | Southwest      | Level 3-4   |
| 2016-06-12 | 31                       | 20                        | cloudy thunderstorms     | Southwest      | Level 3-4   |
| 2016-06-13 | 22                       | 18                        | to heavy rain            | Northeast      | Level 3-4   |
| 2016-06-14 | 24                       | 18                        | thunderstorm to cloudy   |               |            |

Indicator handling considerations

- Statistical time range processing: Since the time range of container utilization data is one year, the existing weather inquiry system can only query the weather conditions after January 2011, so the time range is from January 2011 to December 2016.
- Missing data processing: Since the rainfall and peak sunshine time data of a small number of prefecture-level cities are missing, the county-level city in the area or the nearest prefecture-level city with statistical data is selected instead.
- Processing of imperfect weather information: The weather conditions in a small number of prefecture-level cities have not been counted in January 2011. At this time, county-level cities with complete statistical data in the area are selected instead.

Write crawler programs in python

Use the weather web url+ location to find the location weather query main page based on the target location. The program exports the information of all historical months of the target location to the excel file for further statistical calculation and each month is a sheet. The program directly reads the highest and lowest temperatures of the month, then calculates the average. For the number of sunny days, the number of rainy days, and the number of snowfall days, the program directly reads
the weather conditions of the month, and then accumulates according to the sunny day, rainy days or snowfall days, and outputs the value. For the average sunny time, the average number of rain days, and the average number of snowfall days, the program determines whether there is a corresponding weather every day, and then uses the calculation formula to evaluate. The table of cluster indicators of each city after the program processing is shown in Table 2.

Table 2. Clustering indicator data of various cities in January.

| Location | Month | Average temperature[ °C] | Sunny days[day] | Rainfall [mm] | Rainy days[day] | Snowfall days[day] | Peak sunshine time[h] |
|----------|-------|--------------------------|-----------------|---------------|-----------------|---------------------|----------------------|
| Beijing  | 1     | -1.83                    | 14.8            | 3             | 0               | 1.17                | 2.79                 |
| Tianjin  | 1     | -0.92                    | 12.9            | 3             | 0.25            | 1.17                | 2.81                 |
| Ningde   | 1     | 9.15                     | 3.75            | 80            | 8.58            | 0                   | 2.46                 |
| Fuzhou   | 1     | 9.67                     | 2.58            | 50            | 7.5             | 0                   | 2.46                 |
| Putian   | 1     | 10.52                    | 2.17            | 50            | 5.92            | 0                   | 2.85                 |
| …        | …     | …                        | …               | …             | …               | …                   | …                    |
| Ali      | 1     | -10.53                   | 9.75            | 3             | 0               | 2.33                | 3.61                 |

Determination of the Number of Clusters

Using the weather data of January, use SPSS to generate a cohesive clustering table [6], and then use excel to make a gravel map, as shown in Figure 2. The black box identifies the inflection point of the gravel map as the optimal number of clusters [6], which is 10.

Cluster Analysis Result

Considering different months, the climate types of different cities may have different changes. Therefore, when clustering, according to the monthly index data, cluster analysis is performed according to different months. In this paper, the K-Means algorithm is used to cluster the climate data of different cities in different months. The clusters are clustered into 10 categories, 11 categories, 12 categories, 13 categories, 14 categories and 15 categories for cluster analysis. The clustering results are as follows.

Table 3. KMeans algorithm clustering results of local cities according to weather data.

| Province | City | Month | 15 | 14 | 13 | 12 | 11 | 10 |
|----------|------|-------|----|----|----|----|----|----|
| Beijing  | Beijing | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Tianjin  | Tianjin | 1 | 2 | 1 | 1 | 1 | 1 | 1 |
| Fujian   | Ningde | 1 | 3 | 2 | 2 | 2 | 2 | 2 |
| Fujian   | Fuzhou | 1 | 3 | 2 | 2 | 2 | 2 | 2 |
| …        | …     | …   | … | … | … | … | … | … |
| Xizang   | Ali   | 12 | 1 | 1 | 1 | 1 | 1 | 1 |
The results generated by the K-Means algorithm can be displayed on the map, and the cities with similar colors are one category. The results of national urban cluster analysis based on climate data lay a foundation for the subsequent cluster analysis of railway container handling stations, and also have certain guiding significance for the study of urban climate distribution in China.

Conclusion

According to the research status, it is confirmed that the sun, rain, salt and pollution are the main factors of container corrosion, and the index system of climate data for clustering cities is determined accordingly. Get climate data from the Weather Network (http://www.tianqi.com/) by writing a crawler program. On this basis, using K-Means algorithm, cluster analysis was carried out in 320 major cities across the country. According to different months, clustering results of 10 to 15 categories were generated, and the results were mapped to maps for analysis. It can be seen that the results of clustering basically conform to the actual situation, and have certain guiding significance for the subsequent cluster analysis of container handling stations.

References

[1] Wang xu, Zhu Lingxi. Research on Data Technology of Railway Container Maintenance Management [J]. Railway transportation and economy, 2017, 39 (02):60-64.

[2] Wang Lingli, Zhang Xingchen, Hai tao. Weighted Cluster Analysis of Macroscopic Layout Carriers of Railway Dangerous Goods Handling Stations [J]. Journal of the railway, 2007(06):101-105.

[3] Jin Zhihong, Xu qi, Lan hui. Multidimensional Data Analysis and Data Mining of Container Multimodal Transport [J]. Journal of Jimei University (Natural Science Edition), 2011(04):268-274.

[4] Ning rong. Research on Railway Container Area Transportation Network Based on Clustering Algorithm [D]. Beijing: Beijing Jiaotong University, 2011.

[5] Information on http://www.tianqi.com/

[6] Wang Jingyi. Application of Data Mining Technology in Railway Container Operation and Maintenance Management [D]. Beijing: Beijing Jiaotong University, 2016.

[7] Tan Xiaowei. Classification and Classification of Highway Freight Stations [D]. Xi'an: Chang'an University, 2016.

[8] Xue Wei. Statistical Analysis and SPSS Application (Fourth Edition) [M]. Beijing: Renmin University of China Press, 2014.

[9] Si Shoukui, Sun Xijing. Mathematical Modeling Algorithms and Applications (First Edition) [M]. Beijing: National Defense Industry Press, 2011.