Research on cargo volume prediction and personnel arrangement based on logistics network sorting centers

Hao Zheng *

School of Civil Engineering, Chang'an University, Xi'an, China, 710061

* Corresponding Author Email: 15167720685@163.com

Abstract. In today's increasingly competitive e-commerce logistics industry, various levels of sorting centers on the logistics chain serve as the core nodes and capillaries of the logistics network, bearing the increasingly heavy sorting and transportation functions. Therefore, using the cargo volume of sorting centers to statistically predict future cargo volume and based on this, plan personnel arrangements for each station. This significantly improves the operational efficiency and economic benefits of each station, thereby more reasonably allocating resources, improving service quality, reducing labor waste, improving employee job satisfaction, and achieving the optimization and sustainable development of the logistics industry. This article uses the cargo volume statistics data of 57 sorting centers in a logistics network to analyze the temporal changes of cargo volume at each station on different dates and different time periods on the same day. Starting from the flow of goods within the network and considering the mutual influence of goods volume between stations, machine learning is used to extract the inherent changes in various sorting centers. Regression algorithms based on time series are used to predict the future goods volume for a period. Then, based on the predicted goods volume at each station, a linear optimization is established by considering the constraints of formal and temporary workers to solve the scheduling problem, to achieve a relatively balanced number of personnel for different shifts and minimize the total number of workers and labor costs.

Keywords: Machine Learning, Random Forests, Time Series, Logistics Networks.

1. Introduction

In today's Internet era, logistics, as an important part of the economic cycle, is undertaking an increasingly heavy task of sorting and transferring goods [1]. Freight volume prediction and personnel pre scheduling are two important factors to ensure efficient and timely operation of the logistics industry [2]. Accurate cargo volume prediction helps to allocate human and material resources, improve service efficiency, and scientifically appropriate personnel scheduling can not only meet the work needs of different peak periods, but also minimize the length of employees on duty, making "there are people doing work, people doing work", improving employee satisfaction, reducing costs, and promoting healthy and sustainable development of the industry [3].

A certain logistics network has 57 sorting centers, which need to meet the scheduling and planning problem of formal and temporary workers at SC60 station for the next 30 days. This article establishes a cargo volume prediction model to predict the daily and hourly freight volume for the next 30 days [4]. Then, it considers establishing the corresponding rules of the average freight volume of logistics routes corresponding to a certain origin and destination point and the sorting freight volume of each sorting center [5]. By predicting the relationship between past daily freight volume and the original network, the average freight volume of specific routes for the next 30 days is predicted. Finally, this article establishes a shift-based cargo allocation result and personnel arrangement plan. (Data Sources: http://www.mathorcup.org/)

2. Research on sorting centers and cargo volume

2.1. Visualization and feature extraction of models

To prevent possible missing values in freight volume, Python code is used for visualization processing. The statistical chart shows that there are no missing values in freight volume for all dates, as shown in fig. 1.

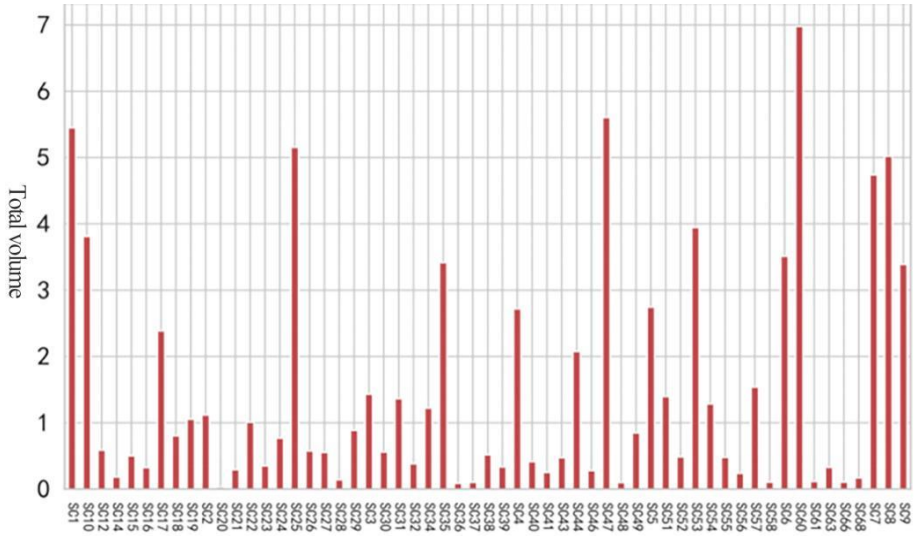

Figure 1. Total cargo volume of different sorting centers

Considering that different sorting centers may have different freight volumes due to factors such as transportation, geographical location, and economy, it is considered to visualize the total cargo volume of each sorting center in the first four months [6]. The specific results are shown in figs 1-3. The sorting modes of each sorting center vary greatly, and it is necessary to establish accurate and targeted prediction models for each sorting center. Therefore, it is possible to consider drawing box plots for each sorting center and analyzing their mathematical and statistical properties. The specific results are shown in fig. 2.

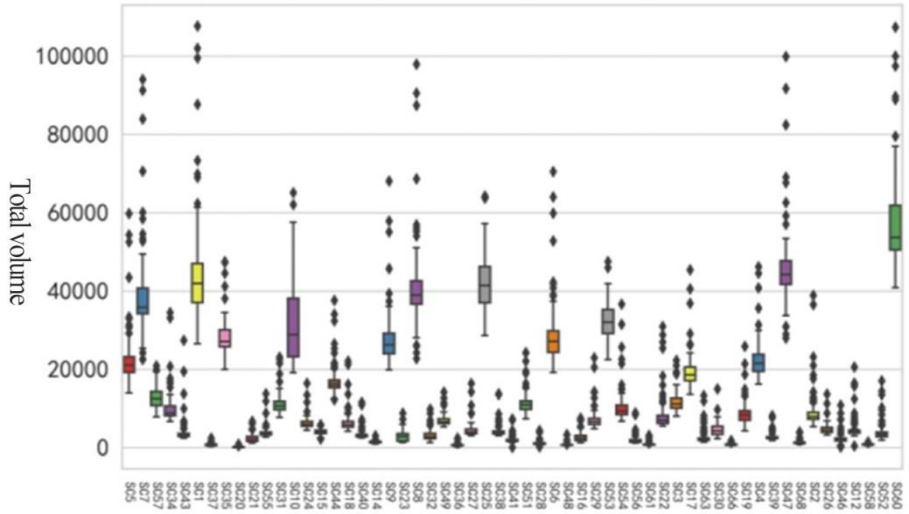

Figure 2. Boxplot of cargo volume differences in different sorting centers

From the box plot, there is no correlation between the mean and variance of each sorting center, indicating that the underlying mechanisms for forming freight volume data for each sorting center are different and not suitable for understanding similar patterns. Therefore, we should consider the trend of the daily freight volume of a single sorting center changing with time coordinates. Therefore, we should consider visualizing the daily freight volume of individual representative sorting centers over time, as shown in figs 3 and 4.

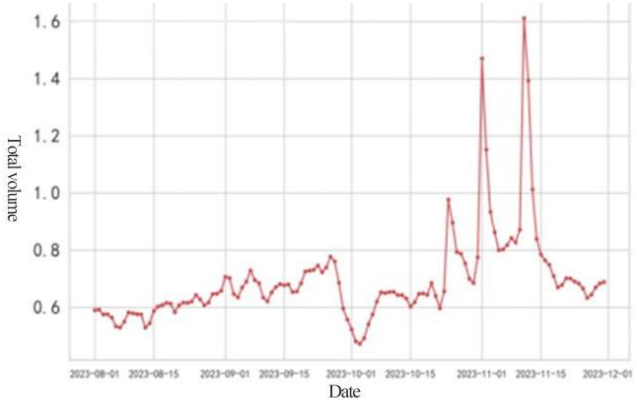

Figure 3. Total freight volume of all sorting centers on different dates

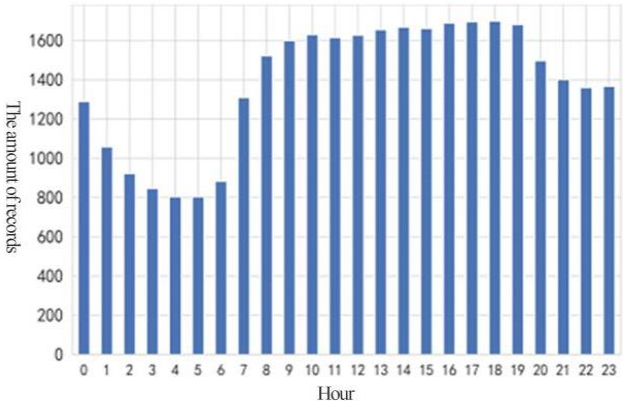

Figure 4. Total freight volume of all sorting centers on different dates

Analyzing the total cargo volume of all sorting centers reveals certain fluctuations and significant local changes. Within the range of August to early October, the total cargo volume maintains a certain linear and slow increase, combined with short-term fluctuations of about a week. This can be explained by common sense [7]. The overall fluctuation may be caused by the season from summer to autumn, cool and comfortable weather, students starting school, and socio-economic operations gradually entering the right track. Cyclical fluctuations may tend to be caused by a high demand for goods orders on a weekday basis, and people tend to rest on weekends and reduce consumption.

Similarly, there is a slight deviation in the hourly statistics of the November records of various sorting centers. If one month is used as the statistical cycle, the total freight volume of each station should have a similar proportion and trend of change. However, when calculating the number of records in Attachment 2, it was found that there were quite a few hours when the number of records was less than the theoretical value of 57 * 30=1710 (records). The results are shown in fig. 4. It is speculated that the possible reason is that the staff between 0:00 and 8:00 did not fully record the freight volume information of all shifts. Therefore, linear interpolation is used here, based on the data correlation between each hour period of the day for linear interpolation. Based on this, a visualization of the total freight volume of sorting centers in November is output, as shown in fig. 5.

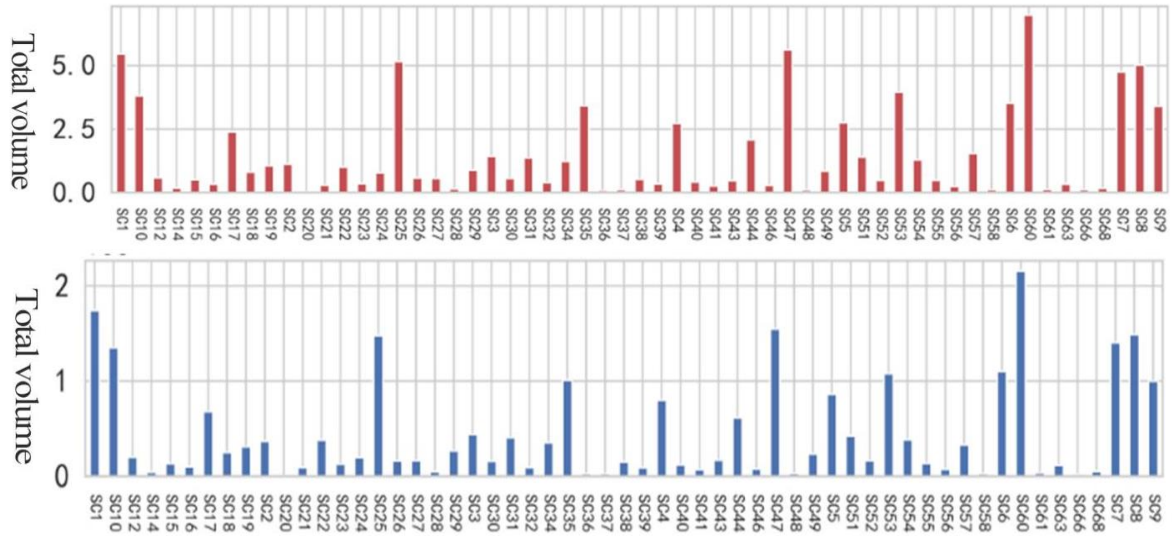

Figure 5. Total freight volume of all sorting centers on different dates

2.2. Modeling unconventional time series characteristics of total cargo volume

The prediction modeling method for conventional data with time as the independent variable is generally time series modeling. However, both ARIMA and SARIMA of time series models have limitations, and the prediction results tend to be smooth for data with large prediction distances.

Considering the implicit impact of historical data on subsequent data, lag value characteristic indicators should be added to unconventional time series predictions, as well as time series characteristic indicators that reflect whether the day is a workday or holiday, and trend indicators that reflect the long-term trend of the dependent variable.

Lag characteristic indicators:

$$Lag_i(t) = Y_{(t-i)} \tag{1}$$

Let Lagi (t) be a lagged characteristic indicator with a lag time of i days for t days as the independent variable, and it can clearly be defined as a function of the dependent variable on day (t-i). Assuming that the regular transportation time of logistics is 3 or 6 days, the freight volume on day (t-6) will inevitably be affected by a nonlinear function on the dependent variable on day t.

$$MovingAverage3(t) = MA3(t) = \frac{1}{3}\sum_{k=1}^{3} Y_{(t-k)} \tag{2}$$

$$MovingAverage6(t) = MA6(t) = \frac{1}{6}\sum_{k=1}^{6} Y_{(t-ki)} \tag{3}$$

In terms of model selection, this article considers various machine learning models, including random forests, neural networks, support vector machines (SVM), and linear regression algorithms. The fitness of different algorithms in this article can be obtained by calculating the fitting attribute scores of the model on the test set and analyzing the residual plot.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i)^2 \tag{4}$$

$$RMSE = \sqrt{MSE} \tag{5}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - y_i\right| \tag{6}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{7}$$

MSE - Mean squared error, RMSE - Root Mean Square Error, MAE - Mean Absolute Error, R^2 Score, N - Number of samples.

After comparing the mean square error (MSE), root mean square error (RMSE), and mean absolute error (MAE) obtained from 57 sorting centers for Random Forest, Neural Network, Support Vector Machine (SVR), and Linear Regression, the accuracy and adaptability of each model were determined. After comparison, it was found that the adaptability and relative error of Random Forest were the smallest, making it suitable to use Random Forest for completion.

2.3. Total cargo volume prediction for sorting centers

The daily cargo volume forecast for some sorting centers is shown in Table 1.

Table 1. Daily cargo volume forecast for some sorting centers

Sorting Center	date	Quantity of goods
SC5	2023/12/2	22367.95
...	...	...
SC5	2023/12/27	21418.62
SC5	2023/12/28	21345.99

3. Research on Predicting Expansion Space in Sorting Centers

3.1. Composition of goods in the sorting center

According to the assumptions and practical knowledge in this article, the task performed by most sorting centers is to sort and organize the goods transported from other sites and the goods delivered by customers (not limited to consumers, but may also be sellers) on this site. After sorting, they are divided into two parts: one is to leave the logistics network and directly deliver the goods offline, and the other is to continue to flow and transfer them to the next sorting center in the logistics network. For the former case, this station is the final pick-up station for the goods, and for the latter case, this station is both the transfer input station for the goods and the transfer output station for the goods. Specifically, corresponding to the final pick-up station, if a certain cargo enters the logistics network at this station, then this station will be the origin and delivery station for that cargo.

3.2. Model construction of sorting centers

Before establishing a prediction model, this article should consider explaining the relationship between known input data and corresponding output data and attempting to interpret the practical significance of the model to make output predictions for another batch of input data.

This article is based on the statistical values of daily freight volume at various sorting centers in the past 90 days and the average freight volume of logistics routes at corresponding origin and destination stations in the past 90 days. If we consider from the perspective of data reality and generation methods, it is precisely the transportation demand corresponding to a specific average cargo volume with a specific origin and destination station that generates the sorting volume for each sorting center [8] [9]. So mathematically speaking, the sorted goods volume at each sorting station can also be obtained from the average cargo volume of each route.

Therefore, this article considers using machine modeling to obtain the correlation between the average cargo volume of the past 90 days and the sorted cargo volume. Since the statistical methods of network organization and logistics flow are the same, using the trained model combined with the average cargo volume data of the next 30 days will inevitably be able to predict the sorted cargo volume for the next 30 days.

$$\begin{aligned}\text{Sorted cargo volume} = {} & \text{The volume of goods originating and delivered on this site} \\ & + \text{The amount of goods transferred and input on this site}\end{aligned} \tag{8}$$

$$\begin{aligned}\text{Sorted cargo volume} = {} & \text{The final pick-up volume of goods on this site} \\ & + \text{The volume of goods transferred and exported from this station}\end{aligned} \tag{9}$$

3.3. Model construction combined with transportation network

When using data from the past 90 days to test the adaptability and accuracy of the model, MSE (mean square error), RMSE (root mean square error), and MAE (mean absolute error) were considered as evaluation indicators. The results showed that random forest is still the most suitable machine learning algorithm for this article.

About Rolling Prediction

Rolling prediction is a time series prediction method that allows models to continuously update their prediction results as new data arrives. This method is very effective for processing real-time data streams because it can adapt to changes in data and provide continuous predictive output.

Let $X(t)$ be the observed value of the time series at time t, and $X(t+1)$ be the predicted value at time t for time (t+1). The process of rolling prediction can be expressed as:

$$X(t+1,t) = f(X(t), X(t-1), X(t-2), \ldots X(t-n+1), theta) \tag{10}$$

Here, f (X) is the prediction function, which depends on the last n observations, and theta is the model parameter.

Over time, when new observations XX become available, the model parameter theta can be updated through optimization methods (such as minimizing prediction errors) to improve future predictive performance. The updated parameters can be represented as theta '.

Therefore, at time (t+1), the updated prediction model can be represented as:

$$X(t+2,t+1) = f(X(t+1), X(t), X(t-1), X(t-2), ...X(t-n+1), theta) \quad (11)$$

This process can be iterated continuously, and as new data arrives, the model continuously adjusts its parameters to provide more accurate predictions.

4. Research on Personnel Attendance Records

This article assumes that there is a difference between regular and temporary workers. Regular workers are fixed at 60 people per sorting center and can withstand a daily sorting cargo volume of 1500 at full load. In sorting centers located at hubs, there is a possibility of insufficient sorting efficiency. Therefore, it is necessary to recruit temporary workers. According to common sense, regular workers are skilled in technology and have higher work efficiency, lower average long-term wages, and extremely high cost-effectiveness [10]. However, due to temporary recruitment, daily wages, and unstable job opportunities, their work efficiency is inevitably low, and their daily wages are relatively high. So, to minimize the total employee attendance shifts (regular worker attendance shifts+temporary worker attendance shifts) while meeting the constraints of sorting efficiency, it is obvious that the maximum number of regular workers should be arranged as much as possible, and temporary workers should be recruited for the excess load. This arrangement has the highest efficiency and the lowest cost, which is the most scientific and reasonable.

4.1. Balance of hourly order volume and allocation of goods shifts

This article is divided into 6 8-hour attendance shifts with repetitive time periods every day, and the predicted results are the predicted sorting freight volume for each hour. Assuming the data is discretized and simplified, all packages that arrive at the sorting station at random times throughout the day are divided into 24 sorting batches, and all goods within an hour are collected at each hour. There are 24 sorting times per day, from 0:00 to 23:00. All goods from 0:00 to 1:00 belong to the 0:00 node, and all goods from 23:00 to 24:00 belong to the 23:00 node.

Let Xt be the predicted value of sorted goods in the t-th hour period, with a value of 0-23. Let set St be the number of all shifts that are currently on duty in the t-th time (A~F, respectively). Let Ci be the sorting task that all personnel in the i-th shift have received at a certain time. For each hour t, execute the following allocation task.

(1) Find the shift set Bt that is currently on duty in the t-th hour

(2) Compare the total number of completed tasks for each shift in the set Bt, select the shift i^* with the lowest received task quantity, and define $i^* = \arg\min(Ci)$.

(3) If all the cargo volume for that hour is allocated to shift A, then Ci=Ci+Xt

From this, this article can obtain the sorting volume of goods allocated to each shift in each sorting center for the next 30 days. The results are shown in Table 2.

Table 2. Forecast of Freight Volume Allocation for Each Shift in Sorting Centers

site	date	Shift	Quantity of goods
SC1	2023/12/1	0--8	19214
SC1	2023/12/1	5--13	9503
SC1	2023/12/1	8--16	3897
SC1	2023/12/1	12--20	5069
SC1	2023/12/1	14--22	6199
SC1	2023/12/1	16--24	5858

4.2. Scheduling optimization problem

Based on the allocation of goods shifts for each hour, the remaining planning problem is only the number of formal and temporary workers corresponding to each shift. The goal of this planning is to minimize the total number of attendance shifts (i.e., one person shifts for each formal or temporary worker who does not attend once) while meeting the restrictions on the number of formal workers, daily work shifts and attendance rates for all employees, and the efficiency requirements of goods sorting

Let Xform (d, i) be the number of formal workers who attended the i-th shift on day d, and let Xtemp (d, i) be the number of temporary workers who attended the i-th shift on day d. The objective function can be written as a minimum expression:

$$Minimize\sum_{d=1}^{30}\sum_{i=1}^{6}(X_{form}(d,i)+X_{temp}(d,i)) \tag{12}$$

$$X_{form}*E_{form}+X_{\text{temp}}*\text{E}_{\text{temp}}\geq P(d,s)\ \forall d\ \text{ə}\ \{1,...,D\},\forall \text{i}\ \text{ə}\ \{1,...,I\} \tag{13}$$

$$\forall d\ \text{ə}\ \{1,...,D\}\ \sum_{i=1}^{6}X_{form}(d,i)\leq R \tag{14}$$

For the constraints, this article aims to ensure that the arranged employees can meet the efficiency requirements of goods sorting. Ensure that the number of formal workers on duty per day does not exceed the total available number, and require the use of formal workers first when both formal and temporary workers are present.

Based on the above cargo shift allocation and personnel optimization arrangement, this article can obtain the required number of formal workers and temporary tools for each sorting center's daily shift. Partial results are shown in Table 3.

Table 3. Plan for allocating formal and temporary workers per shift in the sorting center

Sorting Center	date	Shift	Number of formal workers	Temporary worker count
SC32	2023/12/1	00:00-08:00	6	6
SC32	2023/12/1	05:00-13:00	1	1
SC32	2023/12/1	08:00-16:00	4	4
...	...	...	...	...
SC32	2023/12/7	00:00-08:00	3	3
SC32	2023/12/7	05:00-13:00	1	1
SC32	2023/12/7	08:00-16:00	8	8

5. Conclusion

This article trains the cargo volume prediction model based on historical data to make the prediction results more realistic. This data-driven decision-making approach reduces human intervention and subjective judgment, improving the accuracy and reliability of decision-making. By updating data and models in real-time, enterprises can quickly respond to market changes, optimize resource allocation, and then, based on the predicted cargo volume of personnel scheduling models, achieve refined arrangements for employee attendance plans. By reasonably arranging the shifts of formal and temporary workers, not only does it meet the needs of cargo handling, but it also ensures the work efficiency and quality of employees. At the same time, the model also considers the attendance rate and continuous attendance days of employees, enhancing employee satisfaction and stability. Finally, by optimizing the scheduling plan, the model minimizes the number of people and days while ensuring the handling of goods, thereby reducing labor costs. At the same time, by improving hourly efficiency, work efficiency and overall operational level have been improved. This helps companies maintain a leading position in fierce market competition and achieve sustainable development.

References

[1] Lv Hongyan. Review of Random Forest Algorithm Research [J]. Journal of Hebei Academy of Sciences, 2019, (3): 37 – 41.

[2] Zhu Liping. Application research of enrollment data mining based on decision tree algorithm [J]. Modern Information Technology, 2022, 6 (17): 109 - 112.

[3] Guo Qian Research on multi-layer association rule algorithms and decision tree algorithms in data mining [D]. Dalian Jiaotong University, 2021.

[4] Bai Xiuxian Design and Implementation of a College Enrollment Data Mining and Visualization System Based on Decision Tree Algorithm [D]. Lanzhou University, 2020.

[5] Deng Liantao Research on Logistics Demand Prediction and Development Strategies in Jiangxi Province Based on Machine Learning Combination Models [D]. Fuzhou University, 2021.

[6] Tian Shijie, Zhang Yiming. Overview of Machine Learning Algorithms and Their Applications [J]. Software, 2023, 44 (07): 70 - 75.

[7] Xu X. Automatic classification of transportation modes using smartphone sensors: addressing imbalanced data and enhancing training with focal loss and artificial bee colony algorithm [J]. Journal of Optics, 2024: 1 - 15.

[8] Dong Na, Chang Jianfang, Wu Aiguo. A Random Forest Prediction Method Based on Bayesian Model Combination [J]. Journal of Hunan University (Natural Science Edition), 2019, 46 (02): 123 - 130.

[9] Lee, Sangsuk, and Jooho Kim. "Prediction of nanofiltration and reverse-osmosis-membrane rejection of organic compounds using random forest model." Journal of Environmental Engineering. 2020, 146. (11): 04020127.

[10] Jiao Yangyang, liu Pingzhi, Qi Peixin. Quality Evaluation Method for Settlement Data Matching Based on Grey Correlation Analysis [J]. Journal of Physics: Conference Series, 2022, 2181 (1).