Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

Research on the Volume Prediction of Sorting Center Based on ARIMA-Random Forest

Yuxin Zhao *, Zhaoyang Wang #, Yucheng Chen #, Xin Li #

College of Information Engineering, China Jiliang University College of Modern Science and Technology, Jinhua, China, 321000

* Corresponding author: 15738331387@163.com

#These authors contributed equally.

Abstract. In terms of sorting center volume forecasting, the traditional ARIMA model is difficult to cope with the rapid changes in the market and environment, so an innovative new method based on machine learning is proposed. Then features such as trend, period and lag are constructed for each sorting center to accurately capture the temporal features and improve the prediction accuracy. By comparing various models such as random forest, neural network, support vector machine and linear regression, and comprehensively considering the evaluation indexes such as MSE, MAE and RMSE, the random forest model with optimal performance is finally selected for cargo volume prediction. In the forecasting process, rolling forecasts are used to ensure that the model can be updated in real time and adapt to new market changes. In addition, considering the impact of changes in transport routes on cargo volume prediction, the model is further optimized by introducing transport network information to improve prediction accuracy. This comprehensiveness and flexibility make this method more applicable and effective in practical applications.

Keywords: Sorting Center Cargo; Quantity Forecast; ARIMA; Random Forest.

1. Introduction

Sortation centers in e-commerce logistics networks rely on volume forecasting to optimize resource allocation and reduce costs. Through in-depth analysis of historical data and logistics configurations, sortation centers can accurately forecast daily and hourly volumes, especially after transportation routes have been adjusted, making it critical to update forecast data. Accurate cargo volume prediction helps to rationally allocate shifts and reduce operating costs, thus enhancing the overall efficiency of the logistics network.

Jiang Jun [1] et al. proposed a combined VMD-LR-LSTM model combining variational modal decomposition, linear regression, and long and short-term memory neural networks for the Three Gorges locks cargo throughput prediction problem, which achieves effective prediction of future cargo throughput through decomposing and processing the low-frequency linear part and high-frequency nonlinear part of the original data separately; Wei Lingxiang [2] et al. proposed a time-series prediction method for improving the accuracy of the urban freight volume In order to improve the accuracy of time series forecasting, a forecasting model based on correlation vector machine (RVM) was proposed. The model uses the modelling and solving ideas of RVM to construct the prediction function, and its prediction accuracy is shown to be better than other classical models through example verification.

Previous studies have achieved remarkable results in cargo throughput and urban freight volume forecasting [3], however, in the specific area of sorting center volume forecasting, the traditional ARIMA model is often difficult to cope with due to the rapid changes in the market and environment as well as the complexity of the data characteristics. Therefore, this paper proposes an innovative new method based on machine learning to construct multiple features for each sorting center to improve the prediction accuracy. Afterwards, the optimal random forest model is selected for cargo volume prediction by comparing various stochastic models, and rolling prediction is used to ensure that the

model can be updated in real time and adapt to new market changes. In addition, the model is further optimized by introducing transport network information.

2. ARIMA-Random Forest model for sorting center cargo volume forecasting

Data Source: http://www.mathorcup.org/home/

The data contains daily volumes for 57 sorting centers for the past four months of August, September, October and November; it also contains the average volume of each shipment for these sorting centers for the past 30 days of November; it also provides the average volume of the sorting center's transport routes for the past 90 days; and it also provides the changes in the transport routes for the next 30 days.

Through observation and preliminary data analysis, the data processing involved in this paper includes checking for missing values and transforming the time columns of the given data, as well as linearly interpolating the missing hourly volumes, and constructing features such as lagged features, trend features, periodic features, and moving averages prior to the forecasting of cargo volumes.

2.1. Exploration of the cargo volume data in sorting centers

BP neural network is a multi-layer network with error reverse propagation, which is composed of input layer nodes, hidden layer nodes and output layer nodes. This process has been reduced to an acceptable level of error to the network output, or to a predetermined number of learning times. The network structure is shown in Figure 1.

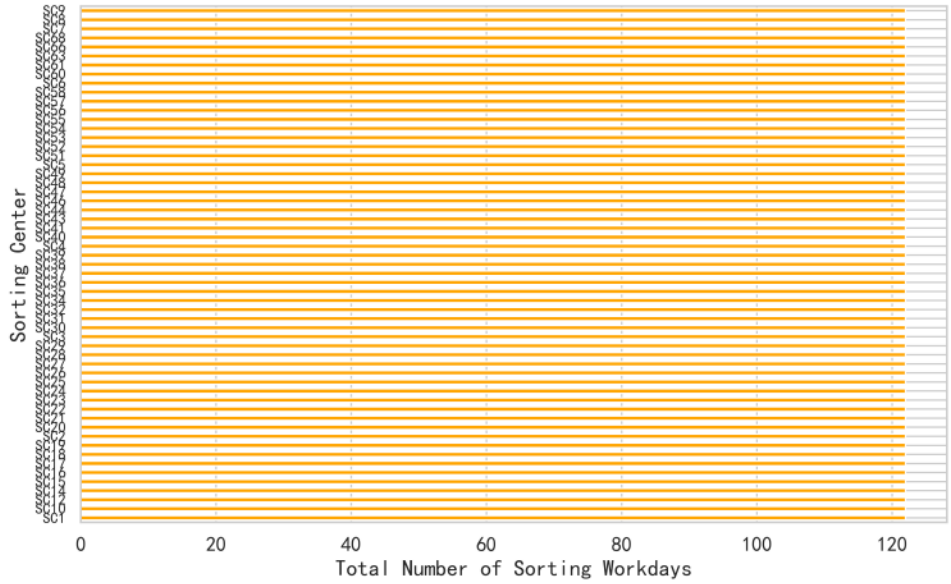

Figure 1. Total Sorting Days of Different Sorting Centers

Cargo volume statistics rely on continuous time series, and data on the number of days at different sorting centers were analyzed and found to be consistently recorded with no missing data.

In addition, factors such as the location of the sorting centers, traffic, population distribution and the degree of urban development varied. In order to intuitively show the impact of these factors on cargo volume, a visual analysis was carried out.

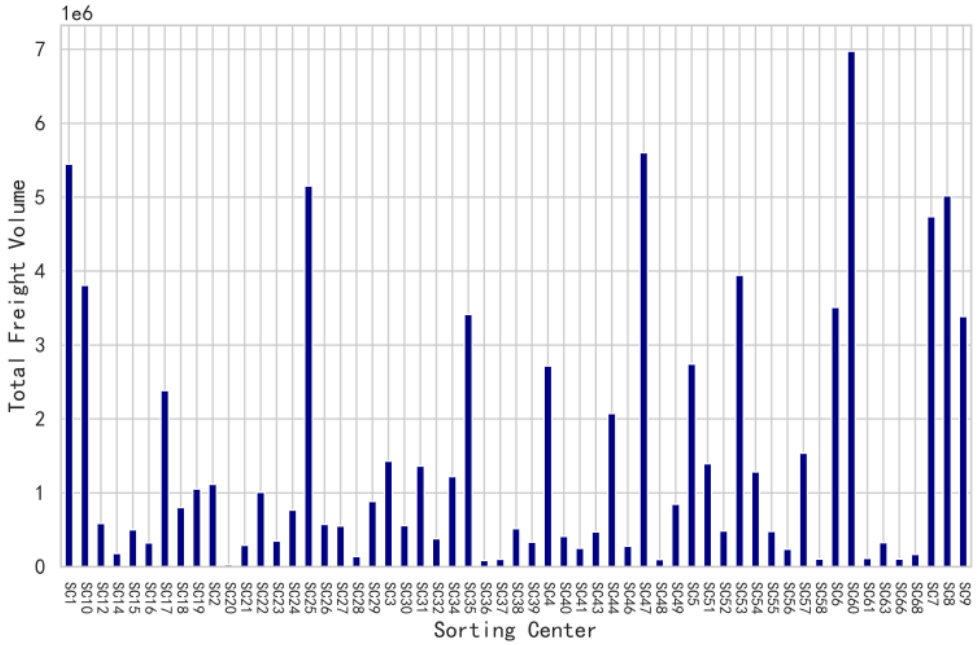

Figure 2. Total Cargo Volume of Different Sorting Centers

As shown in Fig. 2. In Fig. 2, there is a significant difference in the cargo volume of different sorting centers. The difference between less cargo volume such as SC20 and more cargo volume such as SC60 is more than four million. This indicates that each sorting center has a unique pattern, and in order to improve the accuracy of the model, each sorting center needs to be modelled separately to increase the granularity.

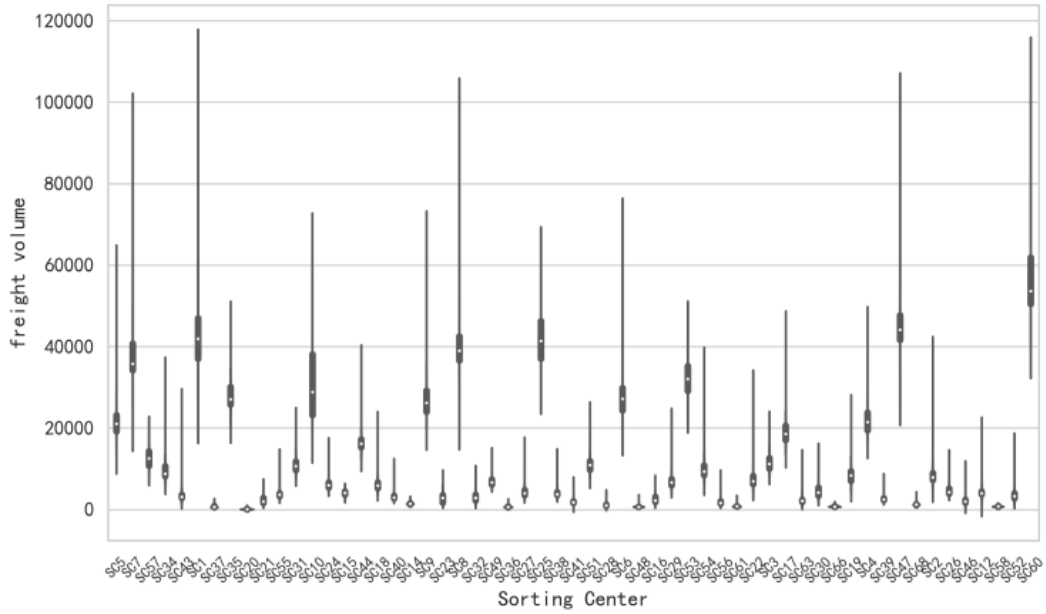

Figure 3. Box Plot of Different Sorting Centers

Figure 3 shows the box-and-line diagram visualization of the difference in shipment volume. In order to improve the machine learning performance and accuracy, the trend of the cargo volume variation in each sorting center needs to be analyzed, as shown in Figure 4, which helps to capture patterns and helps to improve the prediction accuracy of the model.

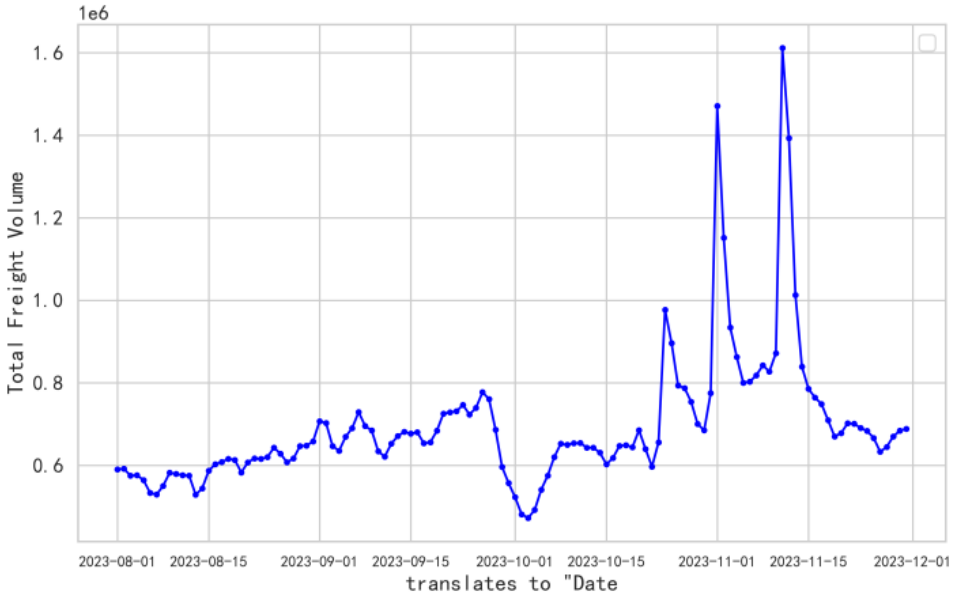

Figure 4. Different total quantity

Figure 4 illustrates the trend in cargo volumes at the 57 sorting centers.

From August to October, cargo volume growth was flat, but there was a significant drop in October followed by a large increase. Into the third week of October, cargo volume began to fluctuate dramatically. By November, the fluctuations were even more dramatic due to the Double 11 campaign, especially peaking on 1 and 11 November, with the largest cargo volume on 11 November. Most of this fluctuation occurs in a small cycle of one week. In order to verify whether this phenomenon is widespread, four sorting centers were randomly selected to view their cargo volume trends, as shown in Figure 5

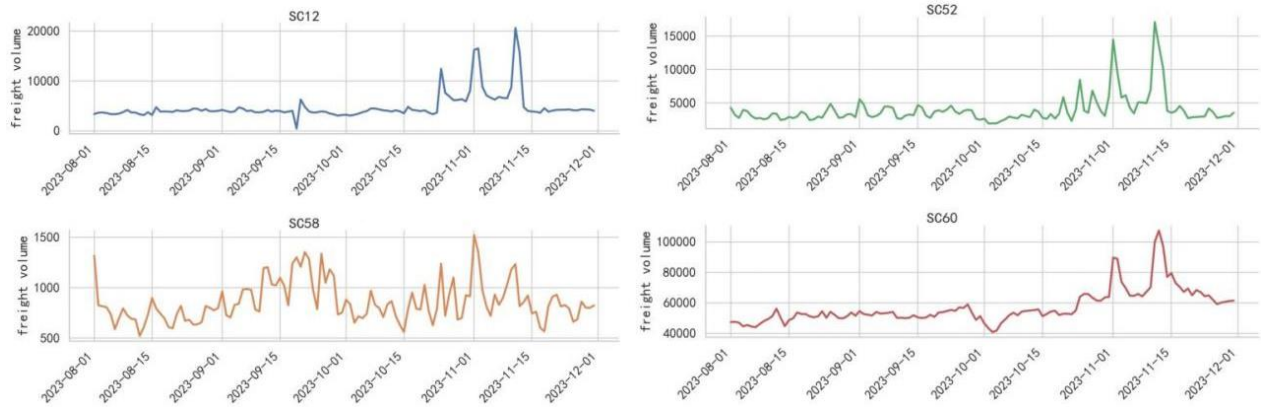

Figure 5. The fluctuation diagram of the cargo volume in the random sampling and sorting center

As can be seen from Figure 5, the quantity of SC12, SC52 and SC60 increased significantly during the Double Eleven event, indicating that the sorting centers SC12, SC52 and SC60 had such a fluctuation, while SC58 did not. First of all, it is necessary to ensure that the hourly supply data is compared with the daily data, and check whether the daily and hourly quantity of 57 sorting centers is consistent. The details are shown in Figure 6 below.

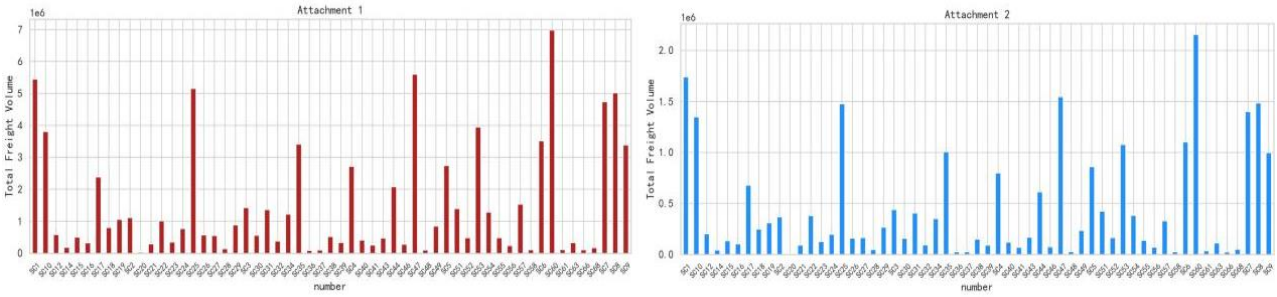

Figure 6. The fluctuation diagram of the cargo volume in the randomly sampled and sorting center

As shown in Figure 6, the daily shipment data and hourly shipment distribution trends are similar, indicating that the hourly data distribution is similar to the daily shipment data.

Therefore, the machine learning model constructed on the daily shipment data may also be applicable to the hourly shipment data. Meanwhile, although the proportion of sorted shipments is similar between the daily shipment center data and the hourly shipment center data, it may differ due to the presence of missing values. For this reason, the hourly shipments for each sorting center were counted for further analysis.

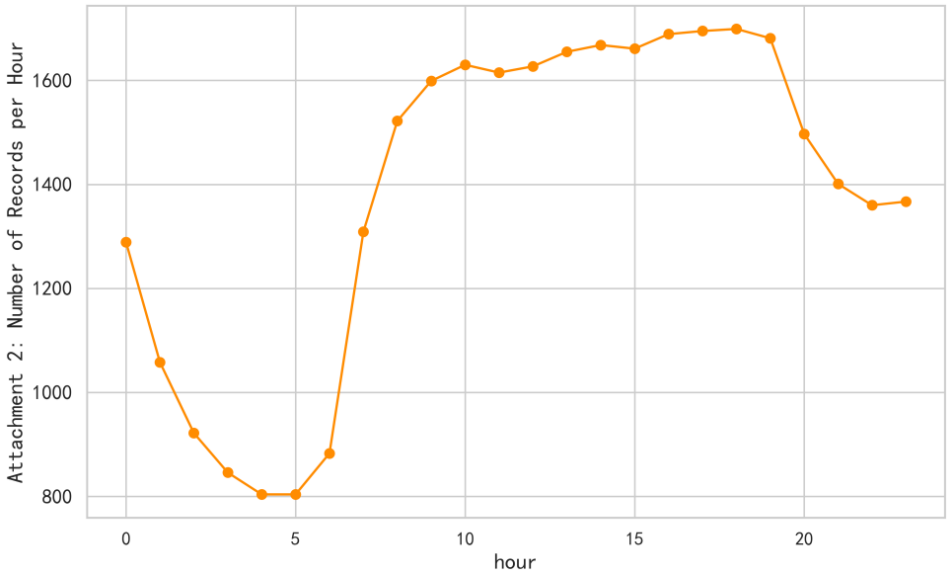

Figure 7. Number of records for different hours

Figure 7 shows that the number of records varies greatly from one time period to another, with better data from 10:00 to 19:00, but many missing values from 0:00 to 6:00, which may be related to operation and management.

The subsequent processing needs to further select linear interpolation for the missing values, as it can retain the original data distribution. After linear interpolation processing, the volume of some sorting centers with very low volume is effectively improved, and the similarity of the ratio of each sorting center with daily volume data rises again, which means that linear interpolation is meaningful, and the accuracy of the model can be further improved by linear interpolation, which can better capture the patterns between different sorting centers. As shown in Figure 8:

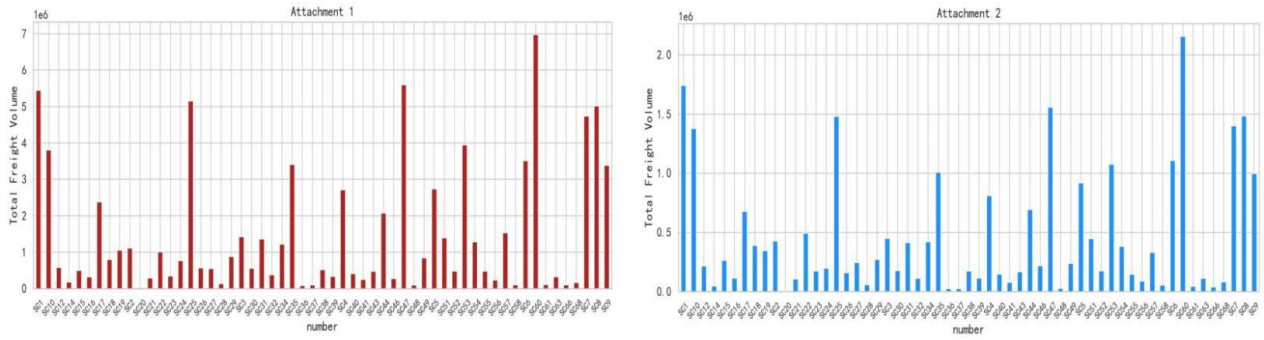

Figure 8. Comparison after linear interpolation of two accessories

2.2. Constructing unconventional time series features

In contemporary logistics systems, sorting center cargo volume forecasting is crucial for resource allocation and process optimization.

Although ARIMA is a commonly used time series forecasting tool, it has limitations in dealing with non-stationary data, seasonal patterns and external influences. In particular, ARIMA may not be able to capture dynamic changes in long-term forecasts and when market conditions change significantly [4]. Its limitations lie in the gradual levelling off of forecasts and the difficulty of fully utilising other relevant influences.

This approach combines manually constructed eigenvalues from historical cargo data, time-series characteristics (e.g., month, day, working day or not), and lagged and moving average cargo volumes. Trend indicators are also considered to reflect patterns of change over time, resulting in a more comprehensive forecast of cargo volumes.

The defined lagged features are:

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

Three-day moving average (3MovingAverage(t))

and six-day moving average (6MovingAverage(t)).

Defined respectively as:

$$\mathrm{3MovingAverage(t) = \frac{1}{3}\sum_{k=0}^{2} y_{t-1-k}} \tag{2}$$

$$6MovingAverage(t) = \frac{1}{3}\sum_{k=0}^{2} y_{t-1-k} \tag{3}$$

2.3. Selection of machine learning models

For model selection, a variety of models that perform well for this problem are considered, including random forests, neural networks, support vector machines, and linear regression [5].

Each of these models will be examined for accuracy and bias based on their scores on the training set, as well as analyzing the residual plots to select the best model to provide accurate predictions for cargo management.

Since different sorting centers are highly variable due to various influencing factors, the model is constructed and trained separately for each sorting center in order to ensure that the model can accurately capture the unique volume trends of each center [6].

In order to achieve this, the features will be constructed using a functionalized approach, evaluated using MAE, RMSE, MSE, and in order to more accurately and easily visualize the overall model performance, the average performance results of MAE, MSE, and RMSE will be derived for each algorithm. Details are shown in Table 1 below:

Table 1. Results of the four algorithms

	Random Forest	Neural Network	SVR	Linear Regression
MAE	0.072376	0.118424	0.101620	0.083501
MSE	0.023476	0.039754	0.024501	0.027811
RMSE	0.167324	0.183841	0.175237	0.168361

Random Forest will significantly outperform the other three algorithms, prior to this, the data has been normalized, so for each model, it is necessary to output its predictions to compare them with the real results and make a residual plot for better analysis. Details are shown in Figure 9 below:

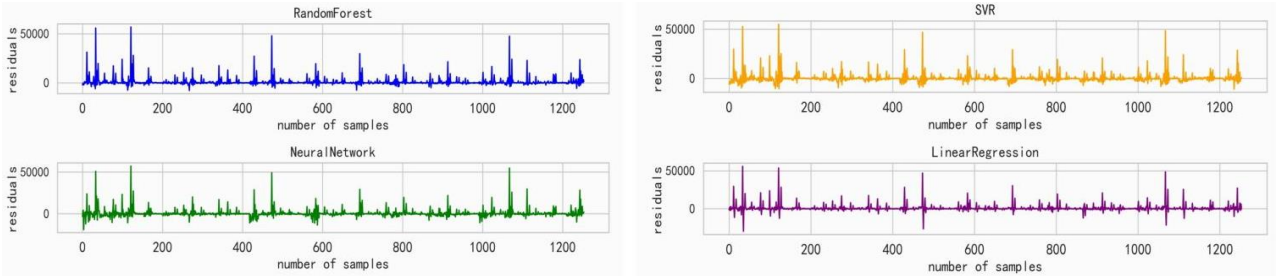

Figure 9. The residual plots of the four algorithms

Observe the volatility of the residual plots, if the residuals are uniformly distributed and there is no significant pattern, it indicates that the predictive model is more effective. Again it can be concluded that Random Forest is the most effective algorithm followed by Linear Regression, Neural Networks, Support Vector Machines in that order.

2.4. Solution by using the random forest model

Finally predict the next 30 days for each sorting center, some results are shown in Table 2 below:

Table 2. Partial predicted results for the next 30 days for each sorting center

Sorting center	date	cargo volume
SC1	2023/12/1	45817.83
SC1	2023/12/2	45917.59
…	…	…

Continue to model the daily inventory data. The results are shown in Table 3 below:

Table 3. forecast each hour for the next 30 days for each sorting center

Sorting center	date	hour	cargo volume
SC34	2023/12/1	15	597.22
SC34	2023/12/1	16	580.86
…	…	…	…

3. Model and eigenvalue solution based on machine learning transport network

An in-depth analysis of the average volume of shipments on transport routes for each sorting center over the past 90 days was carried out, taking into account the volume of goods received and sent and route variations [7].

The transport route characteristics were incorporated into the cargo volume prediction model by extending them. The analysis shows that identifying the transport network is essential to improve the accuracy and robustness of the prediction model [8]. The modelling process extends the features of the random forest model to fully consider the transport network features. This study provides strong support for sorting center operation decision-making.

3.1. Analysis of the transportation network patterns

Constructing valid transport network characteristics is key when predicting the volume of goods at transport network sorting centers.

Average cargo volumes by transport route for each sorting center for the past 90 days were available, but not all 57 sorting centers were involved in the interaction. Therefore, missing data needs to be addressed. The information on daily and hourly sorting center volumes can be combined to identify the sorting centers that are not included in the + average volumes per transport route per sorting center, and the likelihood of missing data can be inferred from this. In order to show the characteristics of the transport network, the Spring layout algorithm is used for plotting [9], as it produces a natural layout, adapts to different network structures and has good interactivity. In plotting this graph, nodes represent sorting centers and edges show distribution volumes. It helps to analyze the specifics of a particular sorting center, As shown in Figure 8:

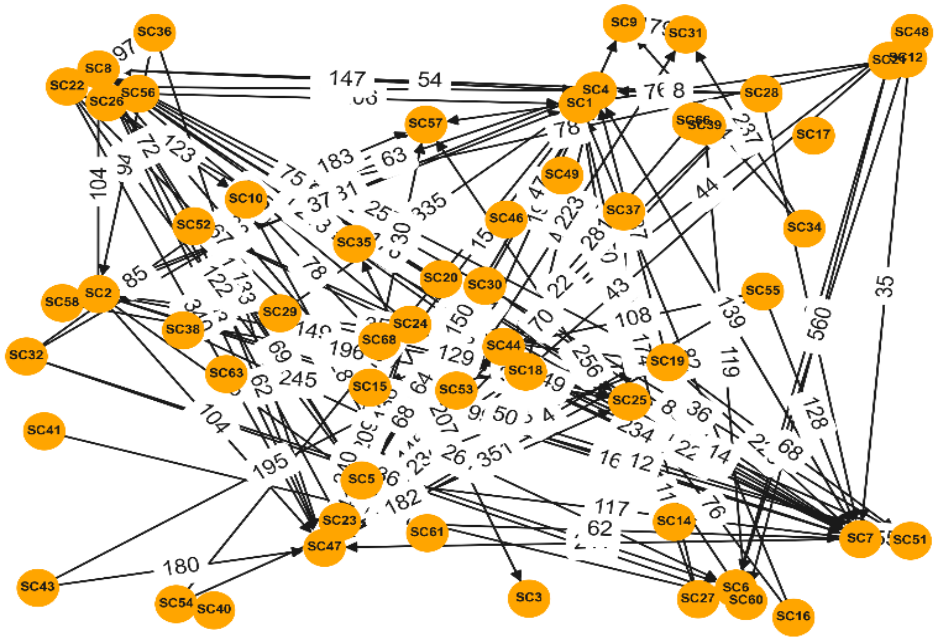

Figure 10. Transportation Network diagram of all sorting centers

3.2. Construct the transport network feature values

The construction of valid characteristics is key when forecasting the volume of goods in the sorting centers of a transport network.

Some of the sorting center data provided for each transport route between sorting centers over the last 90 days is inconsistent with the 57 sorting centers for which daily and hourly volume data is available. This discrepancy may stem from different sorting center functions or data collection errors. Addressing this inconsistency is critical to ensure model accuracy and may involve data population or model adjustments. Considering the data changes over the next 12 months, four features were designed to cope with the impact of transport network changes on cargo volumes, so that transport network features could be constructed efficiently.

1. The average volume of shipments received at this sorting center: total shipments received/total number of days

2. The average volume sent from this sorting center: total volume sent/total number of days

3. This sorting center to receive other sorting center transport volume: total acceptance of other centers volume / total number of days

4. Number of shipments from this sorting center to other sorting centers: /total number of days

Since the exact volume situation of the new transport network is unknown, characteristics can be constructed using the volumes received and sent by the original sorting centers. These features can more comprehensively consider the impact of network changes on cargo volume prediction and improve the accuracy and stability of the prediction mode.

3.3. Model construction of random forest-based transport network and solved using rolling prediction

Similarly, expanding the eigenvalues based on the above proposed solution idea of random forest based cargo volume prediction, the performance between algorithms may produce changes, so again the selection of algorithms and residual map visualization based on the steps of random forest based cargo volume prediction are performed in order to determine whether the solution model should be re-selected or not [10], in turn, The results of the comparison algorithm are shown in Table 4:

Table 4. compares the performance results of the algorithm

	Random Forest	Neural Network	SVR	Linear Regression
MAE	0.083521	0.112587	0.104511	0.071166
MSE	0.024941	0.032204	0.026249	0.018234
RMSE	0.147451	0.191467	0.154231	0.156216

Table 5 shows that Random Forest achieves the lowest scores on MAE, MSE and RMSE and performs optimally Table 5 shows that neural networks outperform support vector machines in problem one, so again algorithm selection is critical. Finally, the results were verified by the second residual plot visualization as shown in Figure 11.

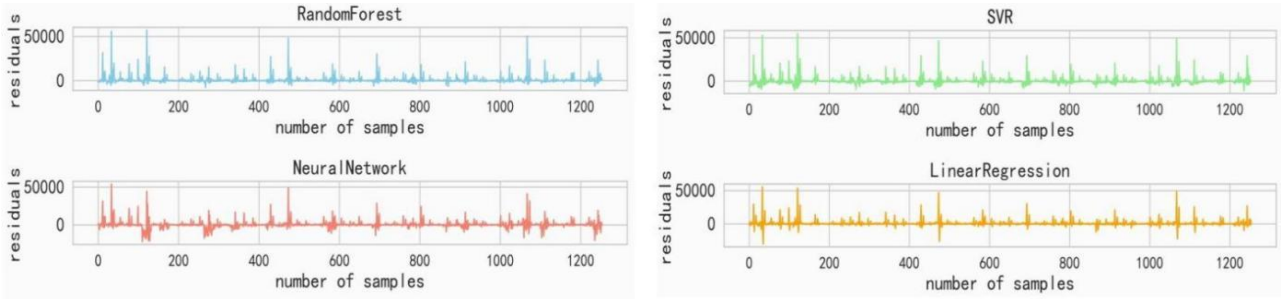

Figure 11. compares the residual maps of the various algorithms after adding features

As shown in Figure 11 above, random forest is still the best choice by observing the fluctuation properties of the residual map.

3.4. The results were calculated by using the rolling prediction

The cargo quantity of each sorting center in the next 30 days is predicted by selecting the random forest combined with the rolling prediction algorithm. Results are presented in Table 5 and Table 6:

Table 5. Daily cargo volume forecast results of some sorting centers

Sorting center	Data	cargo volume
SC5	2023/12/1	22910.82
SC5	2023/12/2	22275
…	…	…

Table 6. Forecast results of daily hourly cargo volume in some sorting centers

Sorting center	Data	hour	cargo volume
SC54	2023/12/1	0	178.55
SC54	2023/12/1	1	121.63
…	…	…	…

4. Conclusion

Firstly, the sorting center volume data is explored and found to be missing in hourly data, then interpolation is used to improve the correlation with daily data so that daily data can be used for uniform analysis, then lagged features, three-day and six-day moving average features are constructed, and combined with the best-performing Random Forest model screened with MSE, MAE, and RMSE in order to improve the accuracy of the model, and for the route change, the for route changes, the transport network eigenvalues are constructed, with four eigenvalues, the average number of shipments received, sent, and the number of shipments received at the current sorting center and sent to the other sorting center, and then the best performing Random Forest model is used again to solve the problem using rolling forecasts.

In the actual prediction of freight volume, will be affected by many factors such as weather changes, traffic conditions, etc., this real-time data integration can help the model to better cope with unexpected events, and further improve the model's adaptability and prediction accuracy. Improving its performance in dynamic environments helps to improve the accuracy of freight volume forecasting and lays the foundation for future staffing and scheduling at each sorting center.

References

[1] JIANG Jun, ZHANG Weiyi, GE Kecheng, TANG Mi, YAN Fanghao, HE Haiting, WEN Yueyu. Cargo throughput prediction at Three Gorges locks based on VMD-LR-LSTM combined model [J]. Water Transport Management, 2023, 45 (11): 6 – 11.

[2] WEI Lingxiang, DONG Jianjun, CHEN Zhilong, WANG Shu. A time series prediction model for urban freight traffic based on correlation vector machine [J]. Journal of Wuhan University of Technology (Transportation Science and Engineering Edition), 2021, 45 (01): 88 - 92.

[3] Zhao Lin, Sun Ningze, Sun Yanbin. Multiple regression analysis of the express delivery industry in Heilongjiang Province [J]. Mall modernization, 2023, (13): 51 - 53.

[4] Zhang Yi. Study and application of time series analysis based on ARIMA-LSTM hybrid model [D]. Changjiang University, 2023.

[5] Wang Qingjie. Study on runoff prediction machine learning models based on feature selection and time-frequency analysis [D]. And Xinjiang Agricultural University, 2022.

[6] Jia Shaobo. Combined prediction method of regional logistics needs for optimal weight determination [J]. Mechanical Design and Manufacturing Engineering, 2020, 19 (08): 113 - 116.

[7] Gupta S, Ali I, Chaudhary S. Multi-target capacity transport: parameter estimation, goodness of fit, and optimization problems [J]. Particle calculation, 2020, 5 (1): 119 - 134.

[8] Jing Li. Research and countermeasures on the development of road Transport Logistics [J]. China Storage and Transportation, 2023, (02): 202 - 203.

[9] Geng Wenjing, Wu Yu. Visualization technology and its current research and application on complex networks [J]. Digital Communications, 2012, 39 (04): 27 - 33.

[10] Wei Lingxiang, Dong Jianjun, Chen Zhilong, et al. Time series prediction model of urban freight volume based on correlation vector machine [J]. Journal of Wuhan University of Technology (Transportation Science and Engineering Edition), 2021, 45 (01): 88 - 92.