Logistics center cargo volume adjustment and personnel optimization based on data analysis

Dongping Sheng, Zhongyuan Ma, Chenqi Zhou, Haidong Feng, Hao Liu, Anbang Zhu, Hun Guo*

Changzhou Institute of Technology, Changzhou, China

ABSTRACT

With the booming development of global e-commerce business, the management efficiency of e-commerce logistics networks has become a focus of attention. In this context, this paper focuses on the prediction of cargo volume and personnel scheduling in sorting centers, aiming to improve the operational efficiency and cost of logistics systems. Three main steps are taken for work one: data preprocessing, stationarity testing, and the establishment of an adaptive hybrid ARIMA-LSTM-XGBOOST weighted model. Data preprocessing includes interpolation of missing values and identification and processing of outliers, supplementing missing values with linear interpolation, and combining JB test and boxplot for outlier detection and processing. Secondly, perform stationarity testing, using the ADF unit root test method to verify the stationarity of the sequence, and use first-order difference to make the sequence stationary. For work two, a complex logistics network topology was established, revealing the structure and layout of the network. Using K-means algorithm to perform clustering analysis on the cargo volume of sorting centers, in order to explore the impact of transportation route changes on cargo volume, and optimize resource allocation to improve sorting efficiency. Using BP neural network for cargo volume prediction, the prediction results and model training state diagram were obtained. For work three, with the premise of ensuring the completion of daily cargo volume processing, an objective function is established to minimize the total number of human days, balance the actual hourly human efficiency per day, hourly human efficiency per site and day, average hourly human efficiency per site and day, and hourly human efficiency variance per site and day. The multi-objective optimization model and genetic algorithm are used to solve the decision variables, aiming to minimize the number of arranged human days as much as possible and pursue the actual hourly human efficiency balance per day. For work four, to ensure the average energy efficiency of the sorting center per hour, scheduling should comply with health and safety standards, ensuring that each shift has sufficient rest intervals to comply with labor health and safety laws and regulations. To achieve this goal, a 0-1 planning model is established to determine constraints such as the total picking quantity not being less than the predicted goods quantity, the maximum daily attendance of formal workers for one shift, the attendance rate of formal workers not exceeding 85%, the continuous attendance days of formal workers not exceeding 7 days and non negative constraints.

Keywords: Stationarity Test, K-Means Algorithm, BP Neural Network, Genetic Algorithm, 0-1 Programming Model

1 INTRODUCTION

In the modern e-commerce environment, the efficiency of logistics networks has become one of the key factors for competitive advantage. The core task of logistics networks is to ensure that every link from suppliers to consumers can be executed quickly and accurately, and the sorting center, as a key node in the logistics network, its efficiency directly affects the operational efficiency and cost of the entire supply chain.

The main function of a sorting center is to process, classify, and forward packages, ensuring that they can be efficiently delivered to the next destination along the correct route and ultimately reach consumers safely. In this process, the management efficiency of the sorting center is crucial, as it not only affects logistics costs, but also affects customer satisfaction and the market performance of the enterprise [1]. In order to improve the operational efficiency of sorting centers, cargo volume prediction has become an indispensable tool. Accurate cargo volume prediction can help logistics managers arrange resources reasonably, including manpower, equipment, and time, thereby optimizing the entire sorting process. In the field of e-commerce logistics, this prediction usually includes two levels: first, based on historical data and logistics network configuration, predict the daily processing volume of goods at each sorting center; The second is to further refine and predict the hourly cargo volume of each sorting center [2]. This detailed prediction is not only helpful for managing daily operations, but also for dealing with sudden increases in workload during promotional activities or holiday peak periods. Personnel scheduling based on predicted data is another key step in optimizing the operation of sorting centers. The personnel in the sorting center usually include two types: formal workers and temporary workers. Regular employees are those who are regularly employed and have high work efficiency and experience; Temporary workers, on the other hand, adjust according to changes in workload, although providing flexibility, usually have lower work efficiency and higher employment costs [3]. By accurately predicting the volume of goods, managers can arrange their personnel structure more reasonably, not only ensuring sufficient labor to handle peak futures volumes, but also reducing unnecessary labor costs when demand is low.

In addition, the application of modern technologies such as big data analysis and machine learning is also changing traditional logistics management models. Through these technologies, sorting centers can more accurately predict cargo volume, adjust human resources more flexibly, further improve efficiency, and reduce operating costs. This not only helps companies reduce internal costs and improve service levels, but also provides consumers with faster and more reliable services, enhancing customer satisfaction and corporate competitiveness. Overall, as the core link of the e-commerce logistics network, the optimization of the management efficiency of the sorting center is an extremely important part of the entire e-commerce supply chain management. Through accurate cargo volume prediction and scientific personnel scheduling, not only can the operational efficiency of sorting centers be improved, but also the operating costs of the entire network can be greatly reduced, ultimately promoting the overall competitiveness of enterprises. To solve such problems, the following four points can be analyzed [4].

Work one: Establishing an e-commerce logistics network sorting center cargo volume prediction model is crucial for resource allocation and operational efficiency in accurately predicting the future cargo volume of each sorting center in the e-commerce logistics network. We need to develop a predictive model to estimate the daily and hourly cargo volume of each sorting center for the next 30 days. This model should help managers effectively allocate resources to cope with predicted changes in cargo volume.

Work two: Adjust the prediction model to cope with potential changes in transportation routes between sorting centers, and the prediction model also needs to be adjusted accordingly to reflect these changes. Explore how to update existing cargo volume prediction models to accurately predict daily and hourly cargo volumes affected by changes in transportation routes over the next 30 days [5].

Work three: Establishing a personnel scheduling model for the sorting center is necessary to optimize the utilization of human resources in the sorting center, especially in the reasonable allocation of the ratio between formal and temporary workers. This model is based on the predicted cargo volume and calculates the number of personnel required for each shift in the next 30 days to ensure timely handling of cargo volume while minimizing the total number of personnel days and maintaining a balance between hourly efficiency.

Work four: Scheduling problem for specific sorting centers (such as SC60). Establish a scheduling model for a specific sorting center SC60, and based on the predicted cargo volume results, plan the shift attendance of each formal and temporary worker within the next 30 days [6]. The model needs to consider multiple constraints, such as a formal attendance rate of no more than 85%, continuous attendance days of no more than 7 days, etc., to ensure that while completing cargo volume processing, it also optimizes the efficiency of human resource utilization and work quality.

2 WORK ANALYSIS AND ASSUMPTION

2.1 Work one

In the process of establishing and solving the model for problem one, data preprocessing is required first. For the data in Attachment 1, use the find function in MATLAB to find missing values. However, for the case where the shipment volume of some sorting centers in Attachment 2 is 0, it is necessary to manually determine these outliers and supplement them with linear interpolation. Then, it is necessary to handle the outliers, use Jarque Bera tests and box plots to identify the outliers, and make manual judgments based on actual situations, such as the quantity of orders on Double Eleven.

After data preprocessing, an adaptive hybrid ARIMA-LSTM-XGBOOST weighted model was established. Firstly, use the ARIMA model to predict the sequence and select appropriate model parameters based on Bayesian information criteria. Secondly, use the LSTM model to capture long-term dependencies in time series data [7]. During this process, the sample data was subjected to maximum/minimum normalization and the input variables of the network were determined. Finally, an XGBoost classification model was established to improve prediction accuracy by constructing an ensemble model of multiple weak learners.

Finally, an adaptive mixed weighting model was established to evaluate the performance of each prediction algorithm and dynamically adjust the weights of each algorithm in the mixed model to improve prediction accuracy. In this process, partial sequences were used as the test set to calculate the prediction errors of each algorithm, and the weights of each algorithm in the mixed model were adjusted based on performance.

Through the above steps, a comprehensive prediction model has been established, which can comprehensively analyze and predict the cargo volume of logistics network sorting centers in the next 30 days, providing more reliable prediction results.

2.2 Work two

Based on the characteristics and data complexity of logistics infrastructure networks, a complex network structure was adopted to establish the topology of logistics transportation routes. This network structure connects various logistics nodes and routes according to the given transportation mode, forming a reliable and efficient logistics network. Firstly, the number of sorting centers and the number of routes connecting them were determined, and the given data was used to calculate the cargo flow between each node, as well as the supply or demand of goods at each node [7].

Next, in order to explore the impact of transportation route changes on the transportation volume of goods in various sorting centers, K-means algorithm was used for clustering analysis. This algorithm automatically classifies the transportation volume of goods in various sorting centers and discovers potential distribution patterns and trends of goods to optimize resource allocation and improve sorting efficiency.

In the process of establishing the K-means clustering model, the cluster center is first initialized, and then the data points are assigned to the nearest cluster center, and the position of the cluster center is continuously updated until convergence. Determine the optimal number of clusters through evaluation indicators or visualization methods, and based on this, perform clustering calculations and visual displays [8].

Subsequently, a BP neural network was introduced for cargo volume prediction. BP neural network, as an unsupervised learning algorithm, can predict cargo volume through historical data. Firstly, a BP neural network model was established, and the processes of forward and backward propagation were introduced. Then, appropriate features were selected based on the feature value selection table, and these features were used to train and optimize the BP neural network. Finally, a daily cargo volume prediction was conducted, and the prediction results were obtained through a neural network model to guide the planning and decision-making of logistics transportation.

2.3 Work three

The purpose of this work is to minimize the total number of person days while ensuring the completion of daily cargo volume, and to strive for a balance in actual hourly human efficiency as much as possible. For this purpose, a multi-objective optimization model was established for this work. In this model, many decision variables and constraints are considered, including the allocation of the number of formal and temporary workers per shift per day, the maximum hourly efficiency of formal and temporary workers, and the predicted cargo volume per sorting center per day.

Firstly, the goal of this work is to minimize the total number of person days. In addition, the balance of actual hourly human efficiency per day was also considered, which is measured by the variance indicator. At the same time, pay attention to the average and variance of hourly human efficiency for a single site and day, to ensure the stability and balance of work efficiency. In addition, it is required that the sorting volume of each shift should not be less than the predicted cargo volume, and constraints such as no more than 60 formal workers and only one shift per person per day are set to ensure the effective utilization of personnel resources and the standardization of work processes.

In order to solve this multi-objective optimization problem, genetic algorithm was chosen for this problem. Genetic algorithm simulates the process of biological evolution, using operations such as selection, crossover, and mutation to evolve generation by generation in order to find or approach the optimal solution [9]. The specific solving process includes steps such as population initialization, individual evaluation, selection operation, crossover operation, mutation operation, replacement operation, and termination condition judgment. Ultimately, through iterative optimization using genetic algorithms, an approximately optimal solution can be obtained, effectively addressing the challenge of daily cargo scheduling, improving human resource utilization efficiency, and ensuring the stability and efficiency of the workflow.

2.4 Work four

This work aims to establish a 0-1 programming model with the goal of minimizing the total number of person days while satisfying a series of constraints. Firstly, the parameters were defined, including the maximum hourly efficiency for formal workers, the maximum hourly efficiency for temporary workers, and the predicted daily shipment volume of SC60. Subsequently, decision variables were introduced, including variables on whether formal and temporary workers are employed, as well as whether formal workers are employed. The objective function is set to minimize the total number of person days.

The constraints cover multiple aspects, including a total picking quantity not less than the predicted quantity, a maximum of 200 formal workers, a maximum of one shift per day for formal workers, a maximum attendance rate of 85% for formal workers, and a maximum of 7 consecutive days for formal workers [9]. At the same time, it also includes non negative constraints as well as constraints that balance the attendance rate of formal workers and hourly efficiency as much as possible.

After establishing a complete multi-objective optimization model, this work chooses to use the multi-objective particle swarm optimization algorithm for solving. Particle swarm optimization is a swarm intelligence algorithm suitable for solving complex optimization problems. Solve the established 0-1 programming model through multi-objective particle swarm optimization algorithm. The algorithm will automatically search for the optimal personnel configuration plan based on the set objective function and constraint conditions. In summary, by establishing a 0-1 programming model and using particle swarm optimization algorithm for solving, problem four can be effectively solved, which is to minimize the total number of person days, achieve effective utilization of human resources, and optimize workflow while meeting a series of constraint conditions.

2.5 Model assumptions

2.5.1 Cargo volume prediction model

(1) Assuming that the logistics network structure remains stable and does not undergo

large-scale changes.

(2) Assuming that the quantity of goods is influenced by factors such as seasonality, holidays, and promotional activities.

(3) Assuming that the trend of changes in cargo volume can be reasonably predicted through time series analysis.

2.5.2 Impact conditions of logistics network changes

(1) Assuming that changes in transportation routes between sorting centers may be influenced by various factors such as geographical conditions, changes in market demand, policies and regulations.

(2) Assuming that changes may lead to adjustments in the flow path of goods, affecting the distribution of cargo volume and transportation demand in the sorting center.

2.5.3 Scheduling Model Assumptions

(1) Assuming that the working conditions and environment of each sorting center are similar, the scheduling model can be universal across different centers.

(2) Assuming that the operational rules and policies of the sorting center are consistent, such as restrictions on continuous working days and maximum attendance rates.

2.5.4 0-1 Planning Model Assumptions

(1) Assuming that the recruitment and management costs for formal and temporary workers are known and remain unchanged during the planning period.

(2) Assuming that the efficiency of goods handling and the operational efficiency of sorting center equipment remain stable and will not change due to personnel adjustments.

3 MODEL ESTABLISHMENT AND ANALYSIS

3.1 Model establishing and solving for work one

3.1.1 Data preprocessing

Work one requires predicting the daily and hourly cargo volume of 57 sorting centers for the next 30 days based on Attachment 1 and Attachment 2. Using the find function in Matlab, missing values are searched for in the given dataset, and it is found that there are no missing values in Attachment 1, but there are a large number of 0 values in some sorting centers in Attachment 2. According to human judgment, it is unreasonable to have 0 values in the shipment volume near Double Eleven [10]. Therefore, it is determined that it is a missing value. As the shipment volume is correlated and cannot be directly deleted, linear interpolation is used to supplement the missing values. The comparison between the interpolation results and the original data is shown in Fig.1.

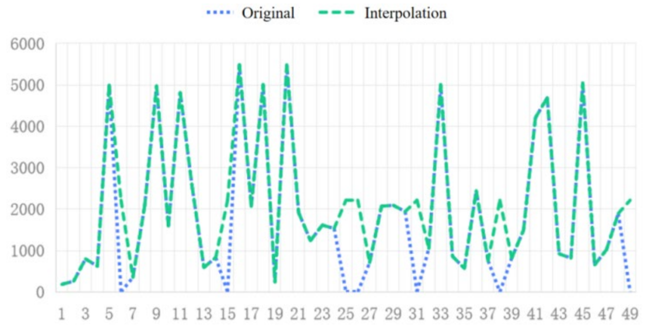


Fig.1 Comparison between raw data and interpolated data

For the handling of outliers, it is necessary to determine the distribution pattern of the dataset given by the work. As the dataset given by the work is relatively large, the Jarque Bera test is used here. JB test is a statistical test method used to test whether data conforms to a normal distribution. It is based on two statistics, namely skewness and kurtosis, to determine the normality of the data by checking whether they are close to the skewness and kurtosis of a normal distribution [11]. JB test is a non parametric testing method that does not require assumptions about data distribution, making it possible to perform normality tests in situations where the data distribution is unclear, making it more flexible. In the JB test, it is assumed that the null hypothesis is that the data comes from a normal distribution. If the data follows a normal distribution, then skewness and kurtosis should be close to 0. The null hypothesis of JB test is that the data comes from a normal distribution, while the alternative hypothesis is that the data does not come from a normal distribution.

The statistical definition of JB test is as follows:

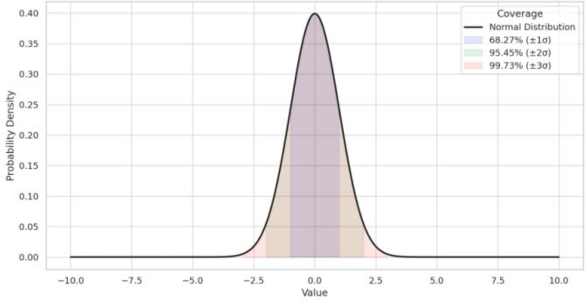
$$JB = \frac{n}{6}\left(S^2 + \frac{1}{4}(K-3)^2\right) \tag{1}$$

Where, n is the sample size, S is the sample skewness, and K is the sample kurtosis.

For data that follows a normal distribution, the 3 σ principle is used to determine outliers. When the given dataset follows a normal distribution, 99.7% of the data falls within the range of three times the standard deviation of the mean. If the data falls outside the range of three times the standard deviation, it is considered an outlier. The formula for determining this outlier is:

$$P(|\ x-\mu\ |> 30) \leq 0.003 \tag{2}$$

Where, x is a data point in the dataset, μ is the average of all data under the corresponding indicator of x, and σ is the standard deviation under the corresponding indicator. The principle diagram of normal distribution is shown in Fig.2, and the judgment results of ten random sorting center box plots are shown in Fig.3.



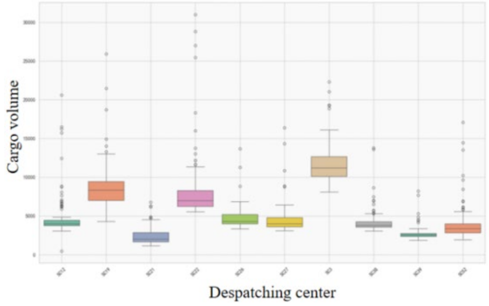


Fig.2 Principle diagram of normal distribution

Fig.3 Sorting Center Box Diagram Judgment Results

For outliers, human judgment should be made based on actual situations. For example, on November 11th, the order volume was 107715, and on the same day it was Double Eleven. Various shopping platforms began to offer a large number of discounts and promotions, resulting in a significant increase in the order volume on that day. Therefore, it was determined by human judgment that it was not an outlier.

3.1.2 Stability test

The method for testing the stationarity of time series generally adopts the ADF unit root detection method. The ADF (Augmented Dickey Fuller) unit root test is a commonly used statistical test method for time series data, which is used to test whether a sequence has a unit root, that is, whether it has non stationarity. The basic principle of ADF test is to test the unit root characteristics of a sequence under the assumption of the existence of unit roots [12].

The original assumption of the ADF test is that the sequence has a unit root, meaning that the sequence is non-stationary. If the null hypothesis cannot be rejected at a significant level (usually 5% or 1%), then the sequence is considered non flat and stable; On the contrary, if the null hypothesis can be rejected at a given level of significance, the sequence is considered stationary. The general judgment formula is:

$$\Delta y_t = y_t - y_{t-1} = \alpha + \beta_t + \delta y_{t-1} + \sum_{i=1}^{p} \xi_i \Delta y_{t-i} + \varepsilon_t \tag{3}$$

Where, Δy_t is the first-order difference of y_t, $y_{t-1}$is the value of the variable at time t and time t-1, α,β, ξ, δ are parameters, t is time, p is the lag order, and ε_t is a white noise sequence.

After performing a first-order difference, a stationary new sequence can be obtained. The differential data is then subjected to the unit root test using Eviews. It can be concluded that the absolute value of the T-statistic is greater than the absolute value of the critical value of the unit root test at each confidence level, and the P-value=0.0001 is less than each confidence level. Therefore, the first-order differential data is a stationary time series.

3.1.3 Establishment of an Adaptive Hybrid Weighted Model

Establishing a weighted hybrid ARIMA-LSTM-XGBOOST model can fully utilize the advantages of the three models, thereby more comprehensively capturing features and patterns in the data. ARIMA models are usually able to capture the long-term trends and periodicity of sequences well, LSTM models can effectively capture and utilize the long-term dependencies in time series data, while XGBOOST models can effectively reduce model bias and variance, improve overall model accuracy by integrating multiple weak learners, and achieve excellent performance on many datasets. Therefore, the combination of the three can better deal with complex data characteristics. This hybrid model can improve the accuracy of prediction, especially when the data contains both linear and nonlinear relationships [13]. The traditional single model is often difficult to fully capture these complex characteristics. By establishing a weighted model, it is possible to comprehensively analyze and predict the cargo volume of logistics network sorting centers in the next 30 days, thereby obtaining more reliable prediction results.

The ARIMA model is a combination of differential operation and ARMA model, denoted as ARIMA(p, d, q). In ARIMA(p, d, q), AR is auto-regressive and p is the number of autoregressive terms; MA is the moving average, q is the number of moving average terms, and d is the number of differences (order) made to make it a stationary sequence.

The ARIMA model can be expressed as:

$$\psi(B) \quad (1-B)^{\mathrm{d}} y_t = \theta(B)\varepsilon_t \tag{4}$$

Where, y_t is the time series of historical observations, d is the order of the difference, ε_t is an independent identically distributed white noise sequence with zero mean and constant variance, and B is the lag factor. B satisfies the following expression:

$$\begin{aligned} B^a y_t &= y_{t-a} \\ \psi(B) &= 1-\psi_1 B-\cdots-\psi_, B^a \\ \theta(B) &= 1-\theta_1 B-\cdots-\theta_q B^a \end{aligned} \tag{5}$$

The key to establishing the ARIMA (p, d, q) model lies in the selection of the three parameters (p, d, q). Here, (BIC) is chosen to select p and q. The Bayesian information criterion can provide a simple approximate logarithmic model evidence, as shown below:

$$BIC = Accuracy(m) - \frac{p}{2}\log N \tag{6}$$

Where, P is the number of parameters and N is the number of data points.

The LSTM (Long Short Term Memory) prediction model is a sequence prediction model based on deep learning, particularly suitable for processing sequence data with long-term dependencies. This model effectively solves the problem of gradient vanishing encountered by traditional recurrent neural networks (RNNs) when processing long sequences by introducing memory units and gating mechanisms. The selection of network input and output variables involves selecting data from three core indicators as sample data to determine the number of network input variables.

XGBoost (eXtreme Gradient Boosting) is an ensemble learning algorithm that is an improved version of gradient boosting trees, particularly suitable for regression and classification problems. XGBoost is widely used in various data mining and machine learning competitions, and has received widespread attention for its efficiency and accuracy.

The XGBoost classification model is an ensemble model that constructs multiple weak learners (usually decision trees). In each iteration, XGBoost adjusts the weight of the current weak learner based on the prediction error of the previous model, thereby reducing the overall loss function of the model. Finally, XGBoost weighted and merged the predictions of multiple weak learners to obtain the final prediction result.

Assuming there is a set of training data, where is the input feature vector and is the corresponding output label. The goal of XGBoost regression is to minimize the loss function, which can be expressed mathematically as follows:

$$\sum_{i=1}^{n} L\left(y_i, \overset{n}{y_i}\right) + \sum_{i=1}^{n} \Omega\,(f_i) \tag{7}$$

Where, L represents the loss function, which is used to measure the error between the predicted value and the true value. $\hat{y}_i$ is the predicted value of the model for the ith sample. $\Omega(f_k)$ is a regularization term used to control the complexity of the model.

XGBoost has excellent predictive performance and can quickly train and predict on large-scale datasets. It can handle high-dimensional features and a large number of samples, and has strong fitting ability for complex data. Support various loss functions and regularization terms, which can be customized according to specific problems. By evaluating the importance of features, it can help select important features and carry out feature engineering. For each prediction algorithm, a partial sequence is used as the test set. The mixing idea of the hybrid algorithm designed in this article is mainly that the better the performance in previous predictions, the higher the weight in future predictions, and the higher the contribution to the predicted values.

For actual observations, they are recorded as:

$$y_i = \{y_1, y_2, y_3, \dots, y_t\} \tag{8}$$

The predicted value of algorithm k is denoted as:

$$\overset{n}{\mathrm{y}}_{\mathrm{k},i} = \left\{\overset{n}{\mathrm{y}}_{\mathrm{k},1}, \overset{n}{\mathrm{y}}_{\mathrm{k},2} \cdots, \overset{n}{\mathrm{y}}_{\mathrm{k},i}\right\} \tag{9}$$

In this work, if the last five days of the sequence are used to calculate the adjusted weights, the prediction error can be recorded as:

$$\mathrm{wmape}_{\mathrm{k,\,id}} = \sum\left|y_i - \mathrm{y}_{\mathrm{k,\,i}}\right| / \sum y_i \tag{10}$$

Total error of algorithm k:

$$\mathrm{wmape}_{\mathrm{k}} = \sum\nolimits_{\mathrm{id}} \mathrm{wmape}_{\mathrm{k,\,id}} \tag{11}$$

In hybrid algorithms, if an algorithm performs well in the test set, the weight will be higher. The weight of algorithm k in the hybrid algorithm can be recorded as:

$$a_{\mathrm{k}} = (1/\mathrm{mape}_{\mathrm{k}}) / \sum\nolimits_{\mathrm{k}} (1/\mathrm{wmape}_{\mathrm{k}}) \tag{12}$$

Where, ω_{k} is the weight of algorithm K in the mixed algorithm. When calculating the mixed prediction value each time, it is necessary to combine the ARIMA algorithm and XGBOOST algorithm with the nonlinear programming algorithm. The calculation formula can be written as follows:

$$\hat{\mathrm{y}}_{\mathrm{i}} = \sum\nolimits_{\mathrm{k}} \omega_{\mathrm{k}} \hat{\mathrm{y}}_{\mathrm{k.\,i}} \tag{13}$$

3.1.4 Model solving

In various logistics sorting centers, XGboost has excellent performance in parallel processing and efficient implementation for the large dataset given to the work. It performs well in both training and prediction stages, and can quickly handle complex problems. And a series of optimization techniques such as gradient boosting and regularization methods are adopted to make the model highly accurate. It can effectively capture complex relationships in data and provide accurate prediction results. If each sorting center has obvious seasonal and periodic characteristics, the ARIMA model has better processing performance for such data, because the ARIMA model can better capture this linear trend and seasonal change, thereby providing more accurate prediction results. The LSTM model is better at handling nonlinear and dynamic changes, especially for data with complex dynamic patterns and long-term dependencies.

Based on the different processing results and ability to capture data features of the three models, they were weighted according to MAPE to ensure the accuracy of the final results. The weighted results are shown in Fig.4 and 5.

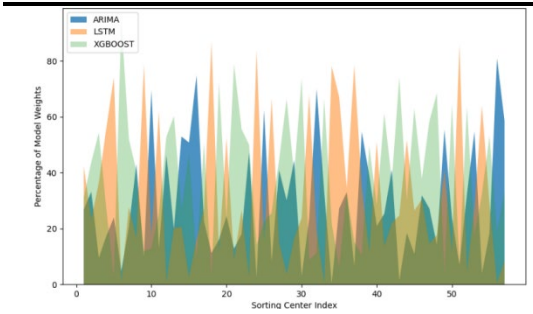


Fig.4 Weights of the mixed model for predicting daily cargo volume in 57 sorting centers

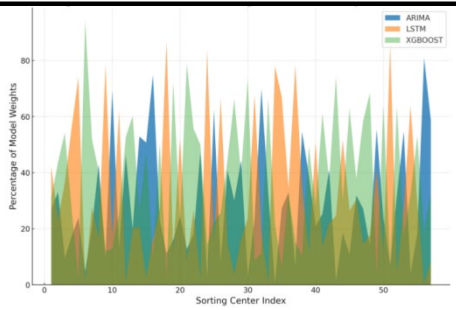


Fig.5 Weights of the mixed model for predicting hourly cargo volume in 57 sorting centers

The predicted daily and hourly cargo volume of 57 logistics centers for the next 30 days is shown in Fig.6 and Fig.7.

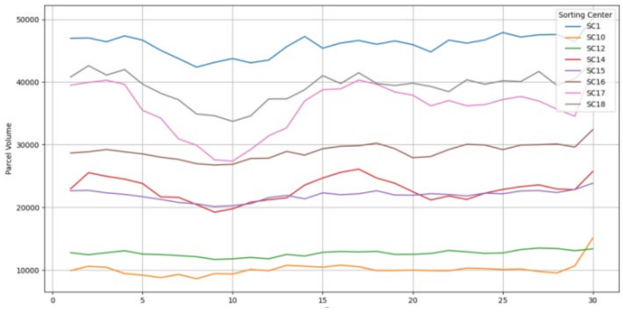


Fig.6 Forecast results of daily cargo volume for the next 30 days

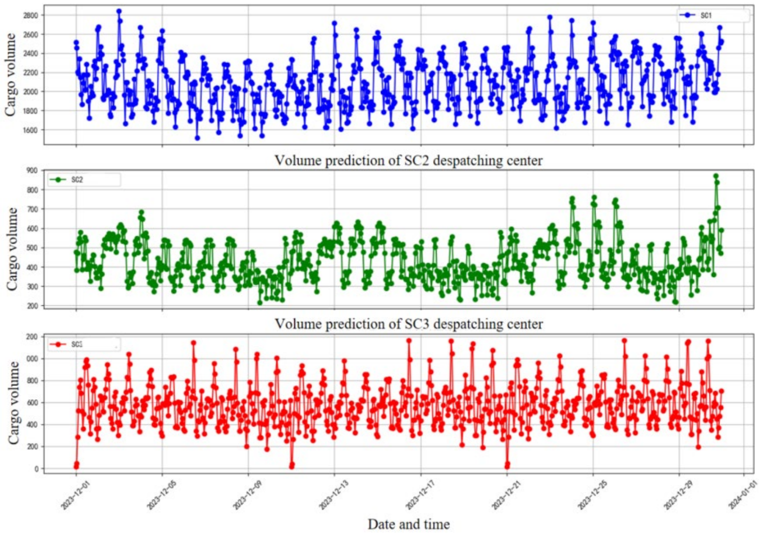


Fig.7 Hourly cargo volume prediction for sorting centers

3.2 Model establishing and solving for work one

3.2.1 Establishment of logistics network topology structure

The basic elements of logistics infrastructure network are logistics nodes and routes. The different ways in which the two are interconnected form different topological structures of logistics networks, from simple to complex: point, line, tree, circle, and network structures.

In this article, a complex network structure is adopted based on the complexity of the data, and various logistics nodes and routes are connected into a network according to the transportation methods provided. In the network, only when logistics business frequently occurs between each node, there are logistics routes between each logistics node to achieve point-to-point connections. Its characteristic is high reliability of network operation, and the addition or subtraction of one or several logistics nodes or routes will not affect the operation of the entire logistics network.

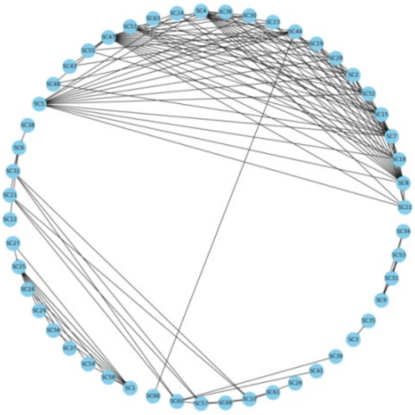


Fig.8 Topological relationship diagram of logistics transportation routes

According to Fig.8, the transportation connection and traffic flow between different sorting centers, as well as the distance and correlation between each node, can be obtained, demonstrating the structure and layout of the logistics transportation network. By analyzing information such as route density and node distribution, one can gain a deeper understanding of the operational mechanism of logistics transportation networks, thereby optimizing transportation plans and improving operational efficiency.

3.2.2 Classification of cargo volume in sorting centers based on K-means algorithm

In order to explore the impact of changes in transportation routes on the transportation volume of goods in 57 sorting centers, the K-means algorithm was adopted to perform cluster analysis on the transportation volume of goods in each sorting center.

The sorting center faces a large amount of goods from other centers and needs to quickly classify and allocate them to appropriate areas for processing. The K-means algorithm can help sorting centers automatically classify goods based on their characteristics, and also help them discover potential distribution patterns and trends of goods, thereby helping sorting centers optimize resource allocation, improve sorting efficiency, and reduce confusion and cross processing troubles.

In the K-means algorithm, the similarity between the dataset and the cluster center continuously updates the position of the cluster center until the cluster center no longer changes. This algorithm requires randomly selecting points as initial cluster centers and continuously updating the position of cluster centers to meet the requirement that all sample points are correctly classified. Although the K-means algorithm is very simple to implement and widely used, it is a hard clustering algorithm, meaning that each sample point can be assigned to a specific cluster.

Assuming the data sample is X, containing m objects, i.e., each containing n attributes of dimensions. The goal of this algorithm is to find the cluster center closest to each object. The Euclidean distance calculation formula between objects and cluster centers is shown in (19).

Among them, X_i represents the i^{th} data object, Y_j represents the j^{th} cluster center, n represents the dimension of the sample data, and X_{it} and Y_{it} correspond to the t^{th} attribute of X_i and Y_j, respectively.

$$D(\mathrm{x_i}, \mathrm{Y_i}) = \sqrt{\textstyle\sum_{t=1}^{n}(\mathrm{X_{it}} - \mathrm{Y_{it}})^2} \tag{14}$$

The K-means clustering model is a commonly used unsupervised learning algorithm used to divide data points into different clusters. The basic steps are as follows:

(1) Initialize cluster center: First, you need to select the initial cluster center, which can randomly hang points in the dataset as the initial center.

(2) Assign data points to the nearest cluster center: For each data point, calculate its distance from each cluster center and assign it to the cluster corresponding to the nearest cluster center.

(3) Update cluster center: For each cluster, calculate the average of all data points and use this average as the new cluster center.

Repeat steps (2) and (3): continue iterating until the cluster center no longer changes or reaches the predetermined number of iterations. Evaluate clustering results: Some indicators such as SSE (Sum of Squared Errors) can be used to evaluate the quality of clustering, and the clustering results can be visualized for intuitive evaluation. Choose the optimal number of clusters: You can try different numbers of clusters and choose the best number of clusters through evaluation indicators or visualization methods.

Firstly, read the data from the data table. Then, identify all relevant nodes and calculate the cargo flow for each node. For each node, calculate the total amount of its shipments and purchases, and select the larger value as the cargo flow for that node. Finally, a dataset containing nodes and their corresponding cargo flow was generated, and clustering calculations were performed on the number of sorting centers based on this dataset. The node and its corresponding cargo flow are shown in Fig.9. This is the visualization result of clustering based on the cargo volume of transportation routes between sorting centers. Each point in the figure represents a sorting center, and different colors and markings represent different clustering groups. The clustering results of transportation routes are shown in Fig.10.

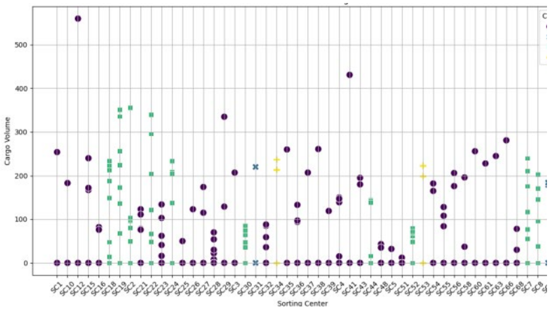


Fig.9 Node and Its Corresponding Freight Flow

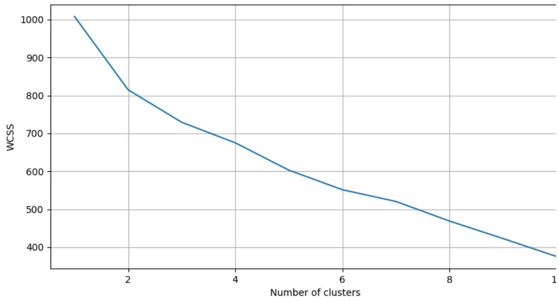


Fig.10 Clustering results of transportation routes

From the elbow method graph, it can be seen that as the number of clusters (k value) increases, the sum of squares within the group (WCSS) gradually decreases and begins to stabilize around k=4. Therefore, k=4 can be chosen as the number of clusters for K-means clustering. Next, k=4 will be used for clustering and a visual display will be created for the clustering results of each sorting center.

3.2.3 Cargo volume prediction based on BP neural network

Neural networks are an important algorithm in machine learning and the foundation for the development of deep learning. It is a system that can self learn, summarize, and generalize through existing data, and can generate intelligent recognition systems through inference, becoming an important component of artificial intelligence technology. With the continuous development of neural network technology, various models have emerged, such as radial basis function neural network, BP neural network, Hopfield neural network, recursive neural network, and AlexNet convolutional neural network.

Among them, BP neural network, as one of the most representative and widely used artificial neural networks, its main advantages include: as long as there are enough hidden layers and hidden nodes, it can approximate any nonlinear mapping relationship; Its learning algorithm belongs to the method of global approximation, which has good generalization ability and fault tolerance. However, the main drawback of BP neural networks is their slow convergence speed, tendency to fall into local minima, and inability to obtain global optimal solutions. In many cases, the number of hidden neurons in BP networks is applicable to the geometric pyramid rule, which means that from the input layer to the output layer, the number of nodes gradually decreases and the shape takes on a pyramid shape. The effectiveness and convergence of the BP neural network algorithm largely depend on the learning efficiency. The use of adaptive learning efficiency algorithms can shorten the training process as much as possible while converging. BP neural network has applications in ballistic simulation, data fusion, radar orbit measurement data anomaly handling, and indoor positioning.

The BP network feature selection table is shown in Table 1.

Table 1 Feature Value Selection Table

Feature	Meaning
The average cargo volume of this sorting center	The average cargo volume of the sorting center given in Attachment 3
Number of upstream sorting centers	Number of upstream sorting centers
Number of downstream sorting centers	Number of downstream sorting centers
Upstream average cargo volume level	Average cargo volume level of upstream sorting centers
Downstream average cargo volume level	Average cargo volume level of downstream sorting centers
Upstream total cargo volume level	The total cargo volume level of the upstream sorting center
Downstream total cargo volume level	The total cargo volume level of downstream sorting centers
Date	The date taken for this data
Time	The time taken for this data

The average cargo volume of the sorting center: The average quantity of goods processed by the specified sorting center during a specific time period. It is an indicator used to measure the cargo handling capacity and efficiency of the sorting center. The average value of the total amount of goods processed within each time unit. By calculating the average cargo volume, information such as the load level, work efficiency, and processing potential of the sorting center can be obtained. This indicator is very important for predicting cargo flow and planning resources.

The daily prediction results are shown in Fig.11 and the hourly prediction is shown in Fig.12.

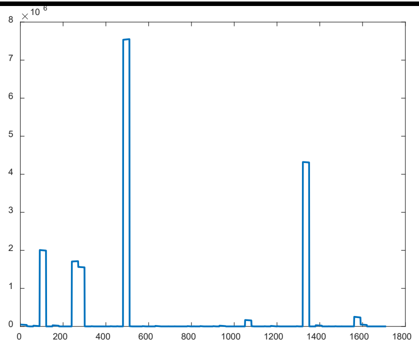


Fig.11 Daily prediction chart based on BP neural network

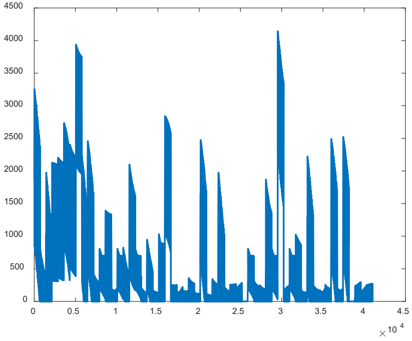


Fig.12 Hourly prediction based on BP neural network

3.3 Model establishing and solving for work one

3.3.1 Establishment of optimization model

For work three, it is required to arrange as few people and days as possible on the basis of completing the daily cargo volume, and to balance the actual hourly human efficiency as much as possible. Establish an optimization model with the total number of person days as the objective function.

(1) The objective function is to minimize the total number of person days:

$$MinZ_1 = \sum_{i=1}^{30}\sum_{j=1}^{6}\left(x_{i,j} + y_{i,j}\right) \tag{15}$$

(2) Balance of actual hourly human efficiency per day (variance indicator):

$$Min\mathrm{Z}_2 = \left(\sum_{\mathrm{i}-1}^{30} D_{\mathrm{i}} - A_{\mathrm{i}}\right)/30 \tag{16}$$

(3) Single site and single day hourly human efficiency:

$$D_i = A_i/8\left(\sum_{j=1}^{6} x_{i,j} + \sum_{j=1}^{6} y_{i,j}\right) \tag{17}$$

(4) Average hourly human efficiency for a single site and day:

$$A_{\mathrm{i}} = \sum_{\mathrm{i}=1}^{30} D_{\mathrm{i}}\,/30 \tag{18}$$

(5) Hourly human efficiency variance for a single site and day:

$$Var = \left(\sum_{i=1}^{30} D_i - A_i\right)/30 \tag{19}$$

(6) The sorting volume of each shift cannot be less than the predicted cargo volume:

$$\mathrm{E}\cdot \mathrm{x}_{\mathrm{i,i}} + F\cdot y_{i,j} \geq C_i \tag{20}$$

(7) No more than 60 formal employees:

$$\sum_k \mathrm{x}_{\mathrm{i,i}} \leq 60 \tag{21}$$

(8) Each person can only attend one shift per day:

$$\sum_j \mathrm{x}_{\mathrm{i}},\mathrm{j},\mathrm{k} = 1 \tag{22}$$

(9) Non negative personnel:

$$\mathrm{x}_{\mathrm{i,j}} \geq 0, \quad y_{i,j} \geq 0 \tag{23}$$

(10) In summary, the multi-objective optimization model established is:

$$MinZ_1 = \sum_{i=1}^{30}\sum_{j=1}^{6}\left(x_{i,j} + y_{i,j}\right)$$

$$\begin{aligned} &= A_j/8\left(\textstyle\sum_{j=1}^{6} x_{i,j,k} + \sum_{j=1}^{6} y_{i,j,k}\right) \\ &= \textstyle\sum_{i=1}^{\infty} D_i\,/30, Var = \sum_{i=1}^{\infty} D_i - A_i/30, x_{i,j,k} \geq 0, y_{i,j} \geq 0 \\ &= \textstyle\sum_{i,j,k} + F \cdot y_{i,j,k} \geq C_i, \sum_j x_{i,j,k} = 1, \sum_i x_{i,j,k} \leq 60 \end{aligned} \tag{24}$$

3.3.2 Model solving

For the established multi-objective optimization model, genetic algorithm is used to find the optimal solution. Genetic Algorithm (GA) is an optimization algorithm based on the theory of biological evolution, used to solve complex optimization problems. It simulates the process of natural selection and genetic mechanisms, and through operations such as selection, crossover, and mutation of individuals in the solution space, evolves generation by generation to produce better solutions, ultimately finding or approaching the optimal solution.

Maximum value issue:

$$F(x) = \begin{cases} f(x) + C_{\min} & \text{if } \mathrm{f(x) + C_{min}} > 0 \\ 0 & \text{if } \mathrm{f(x) + C_{min}} \leq 0 \end{cases} \tag{25}$$

Minimum value problem:

$$F(x) = \begin{cases} C_{\max} - f(x) & \text{if } \mathrm{f(x) < C_{max}} \\ 0 & \text{if } \mathrm{f(x) \geq C_{max}} \end{cases} \tag{26}$$

Perform adaptive scale transformation using roulette wheel selection method:

$$P_i = F_i/\textstyle\sum_{i=1}^{k} F_i \tag{27}$$

Where, Pi represents the probability of the ith individual, Fi represents the fitness of the ith individual, and M represents the number of individuals in the population.

Perform crossover and mutation:

$$\alpha_i' = \begin{cases} 1 - \alpha_i & \text{if } \mathrm{x_i \leq p_m} \\ \alpha_i & \text{if } \mathrm{x_i \leq p_m} \end{cases} \tag{28}$$

Based on stop and constraint conditions:

$$\alpha_i' = \begin{cases} 1 - \alpha_i & \text{if } \mathrm{x_i} \leq \mathrm{p_m} \\ \\ \alpha_i & \text{if } \mathrm{x_i} \leq \mathrm{p_m} \end{cases} \tag{29}$$

Through genetic algorithm, the total number of temporary workers can be obtained as shown in Fig.13, and the specific personnel arrangement for the six time periods of the sorting center is shown in Fig.14.

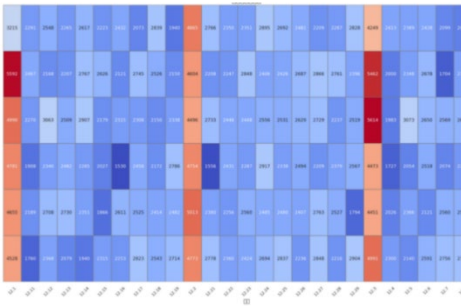


Fig.13 Sum of Temporary Workers

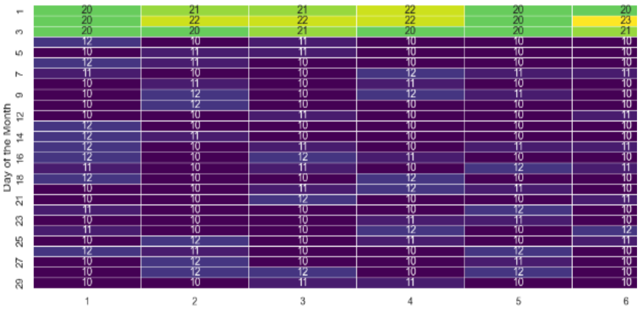


Fig.14 Specific Personnel Arrangement for Sorting Cente

3.4 Model establishing and solving for work one

3.4.1 Establishment of 0-1 optimization model

In this work, establish a 0-1 programming model as follows:

Parameter settings:

E: The highest hourly efficiency for formal workers (25 packages/hour)

F: Maximum hourly labor efficiency for temporary workers (20 packages/hour)

C_i: Predicted shipment volume of SC60 on day i

Decision variables:

$X_{i,j,k}$ is the k^th formal worker of the j^th shift on the i^th day on duty. If working, take 1, otherwise take 0.

$Y_{i,j,k}$: is the k^th temporary worker in the j^th shift on the i^th day working. If working, take 1, otherwise take 0.

W_{ik}: is the k^th formal worker working on the i^th day. Take 1 for work, otherwise take 0.

Objective function: Minimize total person days:

$$Min(Z_3) = \sum_i \sum_j (y_{ij} + x_{ij}) \tag{30}$$

Constraints:

(1) The total sorting quantity cannot be less than the predicted quantity:

$$\sum_{i=1}^{200} x_{i,j,i} \cdot E + y_{i,j} \cdot F \geq C_i \tag{31}$$

(2) The number of formal employees cannot exceed 200:

$$x_{i,j} \leq 200 \tag{32}$$

(3) Regular workers can attend a maximum of one shift per day:

$$\sum_{\mathrm{j}} x_{i,j,t} = 1 \tag{33}$$

(4) The attendance rate of formal workers cannot exceed 85%:

$$\sum_i \sum_i x_{i,j,i} \leq 30.85\% \tag{34}$$

(5) The number of consecutive attendance days for formal employees cannot exceed 7 days:

$$\sum_{\mathrm{i=stret}}^{\mathrm{stret+6}} X_{\mathrm{i,j,k}} \leq 7 \tag{35}$$

(6) Non negative constraint:

$$X_{i,j,k} \geq 0, \quad y_{i,j} \geq 0 \tag{36}$$

(7) Try to balance the attendance rate of formal workers as much as possible:

$$Var(p_j) \tag{37}$$

$$\mathbf{p}_i = \begin{cases} \frac{\sum_{i=1}^{30} \sum_{j=1}^{6} x_{i,j,1}}{200}, \frac{\sum_{i=1}^{30} \sum_{j=1}^{6} x_{i,j,2}}{200} \\ \frac{\sum_{i=1}^{30} \sum_{j=1}^{6} x_{i,j,3}}{200}, \cdots \frac{\sum_{i=1}^{30} \sum_{j=1}^{6} x_{i,j,200}}{200} \end{cases} \tag{38}$$

(8) Strive for a balanced hourly efficiency:

$$\sum_{\alpha=1}^{\infty} V\,ar(E_\alpha) \tag{39}$$

$$E_d = \begin{bmatrix} 25x_{i,j,1} + 20y_{i,j,1}, \\ 25x_{i,j,1} + 20y_{i,j,1} + 25x_{i,d,2} + 20x_{i,d,2}, \\ 25x_{i,d,2} + 20x_{i,d,2} + 25x_{i,j,3} + 20y_{i,j,3}. \\ \cdots, \\ 25x_{i,d,6} + 20x_{i,d,6} \end{bmatrix} \tag{40}$$

The longer part is due to overlapping time periods.

(9) Adaptively arrange work.

Allow for adjusting employee work shifts in special circumstances to adapt to unexpected events or changes.

$$\mathrm{x_{i,j,k}} + \mathrm{x}_{\mathrm{i},j',k} \leq 1 \text{ if } i \neq i' \text{or j} \neq \text{j}' \tag{41}$$

(10) Health and safety standards ensure sufficient rest intervals for each shift to comply with labor health and safety laws and regulations.

$$\mathrm{X}_{i+1,j,k} = 0 \tag{42}$$

If the time interval Xi,j,k=1 is less than the legal requirement.

The multi-objective optimization model obtained from the above is:

$$\mathrm{s.t} = \begin{cases} \sum_{k=1}^{200} \mathrm{x}_{i,j,k}\mathrm{E} + \mathrm{y}_{i,j}\mathrm{F} \geq \mathrm{C}_i \\ \mathrm{x}_{i,j} \leq 200,\ \sum_{\mathrm{j}} \mathrm{x}_{i,j,k} = 1 \\ \sum_{\mathrm{j}}\sum_{\mathrm{i}} \mathrm{x}_{i,j,k} \leq 30\mathrm{x}85\%, \mathrm{x}_{i,j,k} \geq 0 \\ \mathrm{y}_{i,j,k} \geq 0, \mathrm{Var}(\mathrm{p}_i), \mathrm{x_{i,j,k}} + \mathrm{x}_{\mathrm{i}',j',k} \leq 1 \\ \mathrm{x}_{i+1,j,k} = 0,\ \sum_{\mathrm{d}=1}^{30} \mathrm{Var}(\mathrm{E_d}) \end{cases} \tag{43}$$

3.4.2 Model solving

Particle swarm optimization is an optimization method based on swarm intelligence, which simulates the predation behavior of bird flocks. In this algorithm, the potential solution to the work is represented as a group of particles, each representing a candidate solution. These particles fly in search space and update their position and velocity based on their own experience and information from the population. The core idea of particle swarm optimization is to find the optimal solution through information sharing and collaboration between individuals and groups. Each particle will remember the optimal solution they have found (individual extremum), while also tracking the optimal solution found by the entire population (global optimal solution). In each iteration, particles update their velocity and position based on their own velocity and position, as well as information on individual extremum and global optimal solution. This involves formulas for the position and distance of particles.

$$\begin{aligned} &z_{\mathrm{ij}} = z_{\mathrm{ij}} + d_{1\mathrm{ij}} \times rand() \times \left(pbest_{ij} - x_{ij}\right) + d_{2\mathrm{ij}} \times rand() \times \left(pbest_{ij} - x_{ij}\right) \\ &x_{ij} = x_{ij} + z_{\mathrm{ij}} \end{aligned} \tag{44}$$

Where, rand() is the random number between 0 to 1.

$$\begin{gathered} z_{\mathrm{ij}} = \omega \times z_{\mathrm{ij}} + d_{\mathrm{iij}} \times rand() \times \left(pbest_{ij} - x_{ij}\right) \\ +d_{2\mathrm{ij}} \times rand() \times \left(pbest_{ij} - x_{ij}\right) \end{gathered} \tag{45}$$

Where, the inertia factor ω has a non-negative value. The larger the value, the stronger the ability to find the optimal solution, and the weaker the ability to locally find the optimal solution. Conversely, the opposite is true. We adopt a linear decreasing weight strategy here.

$$\omega^{(t)} = (\omega_{ini} - \omega_{end})(G_k - \mathrm{g})/G_k + \omega_{end} \tag{46}$$

Where, G_k is the maximum number of iterations, ω_{ini} is the initial inertia weight, and ω_{end} is the inertia weight at the maximum number of iterations.

When initializing the particle swarm, a random set of initial particles will be generated as possible solutions. The initial positions and velocities of these particles will be dynamically adjusted in the subsequent optimization process based on their own historical best experience and the best experience of the entire particle swarm. To evaluate the superiority or inferiority of each particle, it is necessary to define a fitness function. This function will evaluate the quality of the personnel configuration scheme represented by each particle based on the goal you set, which is to minimize the weighted sum of total person days and efficiency deviation.

The position and velocity updates of particles follow the standard rules of PSO, which combine the best experience of the particles themselves with the best experience of the entire population. By continuously updating the position and velocity of particles, the PSO algorithm can guide the particle swarm to move towards a better solution space region. Finally, the PSO algorithm gradually improves the solution of the particle swarm through multiple iterations until the preset stopping conditions are met, such as reaching the maximum number of iterations or the quality of the solution no longer significantly improves. In this way, an optimized personnel configuration plan can be found through the PSO algorithm to minimize the total number of person days and efficiency deviation.

4 CONCLUSION

In the modeling process of work one in this article, three models, ARLMA, LSTM, and XGBOOST, were used to predict data for the next 30 days. The following conclusions have been drawn through research and analysis:

The ARLMA model can dynamically adjust weights based on the characteristics of input data, improving the accuracy and adaptability of the model; The LSTM model can effectively capture and model long-term dependencies in sequence data, which is very useful for tasks that require understanding contextual information. Moreover, the LSTM model can store and update information through memory units, and has strong long-term memory ability, which can capture and learn important features in sequences. The XGBoost model has excellent performance, can handle large-scale datasets, and can achieve fast speed in both training and prediction stages. It can handle various types of features, including numerical and categorical features, and can automatically handle works such as missing values and outliers, reducing the burden of feature engineering.

A BP neural network model was established to predict the cargo volume after the change of the second work route. The BP neural network model can approximate any complex nonlinear function relationship, has strong pattern recognition and fitting capabilities, and each neuron can be computed in parallel, improving computational efficiency. It has certain advantages in processing large-scale datasets, and can also use back propagation algorithm to train the network. By continuously adjusting weights and biases, the predictive performance of the model can be continuously optimized.

In the third work, in order to achieve a reasonable arrangement of attendance, this article establishes a genetic algorithm optimization model. Genetic algorithms have strong ability to handle complex problems, especially suitable for problems that cannot be solved through mathematical modeling, such as combinatorial optimization, path planning, etc. Under multiple constraints, it can effectively avoid getting stuck in local optima, thus conducting a global search in the entire search space and achieving scheduling optimization for attendance personnel.

The fourth work adopts the 0-1 programming model, which can be applied to various decision-making problems, such as resource allocation, path selection, scheduling problems, etc. Since variables can only be taken as 0 or 1, the 0-1 programming model can ensure that the obtained solution is feasible and meets practical application needs.

5 ACKNOWLEDGEMENTS

This work was supported by ministry of education industry-university cooperative education project (Grant No.: 231106441092432) and special research project on teaching reform (Grant No.: 30120300100-23-yb-jgkt03)

REFERENCES

[1] Geng, Y., & Ju, S. D. (2005). Research on the Topology Structure and Governance Model of Logistics Infrastructure Network. *2005 National Doctoral Academic Forum (Transportation Engineering)*.

[2] Luo, J. X., & Ding, B. X. (2021). Research on the Analysis and Prediction Method of CPI Economic Indicators Based on ARMA and LSTM. *Journal of Mount Huangshan University*, 23(05): 14-18.

[3] Zhang, P. L. (2021). Research on optimization of storage capacity of fast-moving consumer goods based on order prediction. *North China Electric Power University*.

[4] Yan, N. (2022). Research on volume prediction and route optimization of H company's container dumping transportation. *Guizhou University*.

[5] Shen, Q. M. (2023). Empirical Comparative Study of ARIMA, LSTM and their combination models. *Suzhou University*.

[6] Xu, B., & Feng, Z. (2023). Study on electronic product assembly line balance based on integer programming and simulation. *Mechanical & Electrical Engineering Technology*, 1-8.

[7] Sun, S. Y. (2023). Research on Cargo Loading and Unloading Point Identification and Logistics Center Site Selection Based on Trajectory Data. *Dalian Maritime University*.

[8] Pan, Q. Y., & Yang, J. Q. (2023). Research on Classification of Interior Materials Based on Improved K-means Clustering Algorithm. *Proceedings of the 2023 World Congress on Transportation*.

[9] Zhang, X. G., & Luo R. (2024). Analysis of Traditional Chinese Medicine Constitution Prediction Based on ARIMA Time Series Model. *Asia Pacific Traditional Medicine*, 20(04): 156-162.

[10] Shen, L., Chen, X., & Tao, W. B. (2024). Application of BP neural network based on genetic algorithm in settlement prediction of light subgrade. *Journal of Guangxi University of Science and Technology*, 2: 32-39.

[11] Liu, H., Yuan, X. J., & Lv, C. R. (2024). LSTM neural network based bridge displacement missing data reconstruction method. *Journal of Wuhan University of Technology(Transportation Science & Engineering)*, 1-8.

[12] Xie, Z. X., & Zhang, Z. Y. (2024). Research on early warning of unsafe behavior among miners based on GA optimized BP neural network. *Mining Technology*, 24(02): 137-142.

[13] Che, G. Q. (2020). Heuristic mixed integer linear programming for flight conflict resolution. *Zhejiang University of Technology*.