A Study on Cargo Volume Prediction and Staff Scheduling Based on Support Vector Regression Model and Integer Planning

Yishan Wu *

Yanjing Medical College, Capital Medical University, Beijing, China, 101300

* Corresponding Author Email: w15022378142@163.com

Abstract. In pursuit of enhancing the management efficiency and reducing operational costs within the logistics network sorting centers, this research constructed a Particle Swarm Optimization (PSO)-Supported Vector Regression (SVR) model coupled with an Integer Programming staff scheduling model. Utilizing tools such as MATLAB, Excel, and Python, the study conducted processing and analysis of the cargo volume data, revealing the periodic trends of the cargo volume. After eliminating outliers, the optimized SVR model was employed for forecasting, and weights were established to adjust features and mitigate the impact of non-interactive sorting centers. The optimal staff allocation for each sorting center was determined through Integer Programming. This research provides an efficient set of strategies for cargo volume prediction and workforce scheduling, contributing to the optimization of cost-effectiveness in logistics management.

Keywords: Support Vector Regression, Particle Swarm Algorithms, Scheduling Problems, Integer Planning.

1. Introduction

With the rapid development of e-commerce logistics networks, accurate forecasting of sorting center volume and optimization of staff scheduling have become crucial. In addressing the issue of freight volume prediction, existing solutions include Li Hongjuan et al. [1] use seasonal decomposition to analyze national freight data according to seasonal fluctuation patterns and established the SAO-LSSVM-MA model for prediction based on the characteristics of the decomposed data, and Huang Juan et al. [2] employ a combined forecast using linear regression and trend prediction models. Regarding staff scheduling, current solutions involve establishing set-partitioning scheduling models aimed at minimizing shift costs [3]. This paper utilizes a Support Vector Regression (SVR) model to forecast the operational volume of data from sorting centers and based on these predictions, employs integer programming algorithms to arrange staff attendance schedules. (Source data: https://pan.baidu.com/s/1rXpG4vQSvtUIFIZcSCe_PQ?pwd=9694).

2. The Support vector machine model for the amount of cargo forecasting

2.1. The daily hourly distribution of volume in sortation centers

Initially, this paper employs the Python programming language to conduct a thorough analysis of logistics data. By categorizing the data on a monthly basis, an overall upward trend in cargo volumes during August, September, and October is shown in fig.1. Notably, the total cargo volume in November is significantly higher than the preceding three months, peaking during the "Double Eleven" shopping festival. This phenomenon indicates that promotional activities by e-commerce platforms have a substantial impact on the sorting center's cargo volume. To ensure the accuracy of the model, this paper adopts a strategy of excluding outliers, specifically removing records with cargo volumes exceeding 70,000.

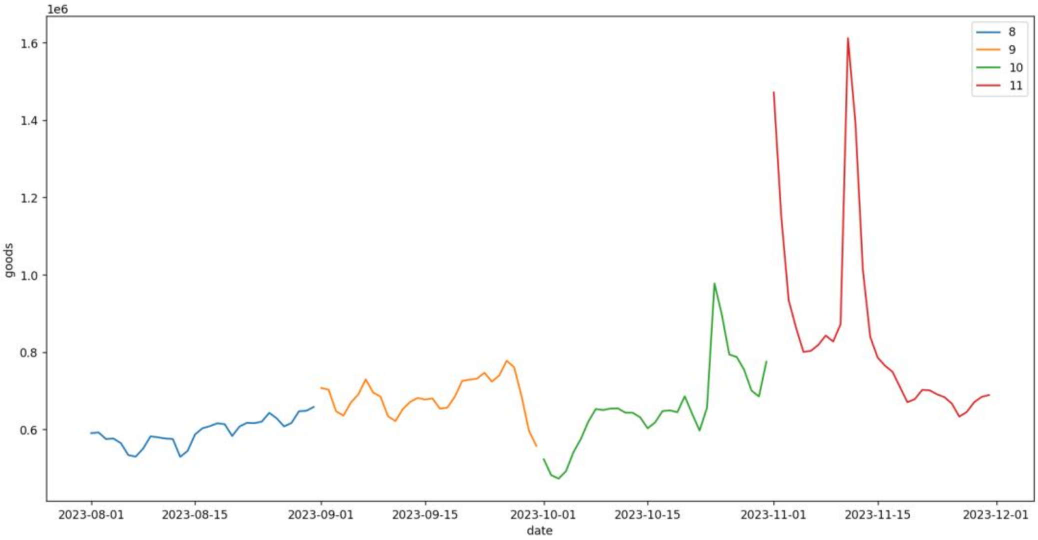

Fig 1. Daily volume from August to November.

Subsequently, this paper categorizes 33,281 data entries on an hourly basis and illustrates the results in fig.2. The analysis indicates that the overall trend is consistent with the data from November. Within the time frame of 0-23 hours, there is a significant temporal disparity in the distribution of cargo volume, with a notable increase in logistics cargo volume during nighttime compared to early morning hours. Similarly, to mitigate the potential impact of outliers on model fitting, this study resolves to remove data records with cargo volumes exceeding 70,000 as well as those with zero cargo.

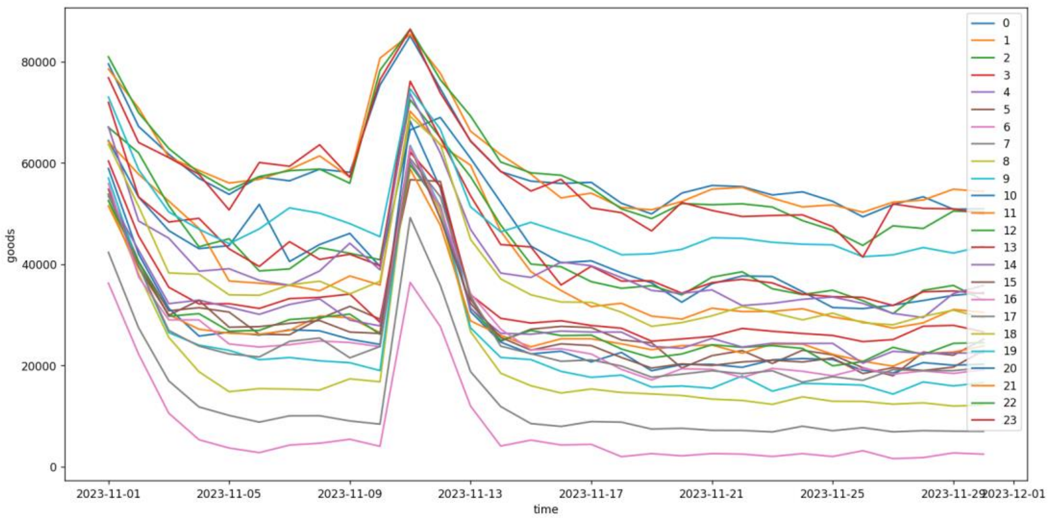

Fig 2. Hourly volume from November 1st to November 30th.

2.2. The structure of Support Vector Regression

This study adopts the Support Vector Machine (SVM) as the principal modeling tool. The SVM is a pattern recognition method grounded in statistical learning theory. Its core concept revolves around finding a hyperplane that correctly separates the two classes of data points to the greatest extent possible, while also maximizing the margin between these separated classes [4]. Compared to traditional regression methods which consider a prediction correct only if the regression f(x) equals y exactly, Support Vector Regression (SVR) deems a prediction accurate if the deviation between f(x) and y does not exceed a specified value, thus eliminating the need for loss computation [5].

When employing the SVM for regression tasks, the kernel parameter γ is used to map sample data into a higher-dimensional space, fitting the data $\{x_i, y_i\}$ with the function $f(x) = wx + b$, where i =

1, …, k, $x_i \in R^d$, and $y_i \in R$. To accommodate allowable fitting errors, a slack variable $\zeta_i \geq 0$ is introduced, permitting a minimal number of samples to be misclassified. The degree of this error is controlled by the parameter c [6]. The optimization problem is described as follows

$$min \frac{1}{2}||w||^2 + c\sum_{i=1}^{n} \zeta_i \tag{1}$$

$$s.t. \begin{cases} g_i(w^T t_i + b) \geq 1 - \zeta_i \\ \zeta_i \geq 0, i = 1,2, \ldots\ , n \end{cases} \tag{2}$$

Then seeks to find the optimal values for w* and b*. This study treats data from each sorting center and dates as input features, with cargo volume as the output target. The rbf kernel function and linear kernel function are used to fit and predict sample data. To measure the model's prediction accuracy, the Mean Absolute Percentage Error (MAE) and the R^2 are defined, which represents the proportion of variance explained by the model. Yi is the actual value of the i-th learning sample during fitting, and Yi' is the predicted value of the i-th learning sample.

$$MAE = \left(\frac{\sum_{i=1}^{n} \left| \frac{Y_i - Y_i'}{Y_i} \right|}{n} \right) * 100 \tag{3}$$

$$R^2 = 1 - \left(\sum_{i=1}^{n} (y - \hat{y}) \sum_{i=1}^{n} (y - \bar{y}) \right) \tag{4}$$

In predicting the daily volume of goods at sorting centers, we employ the minimal linear kernel function's results as the projected figures for the upcoming month. When forecasting the hourly daily volume, considering the extensive dataset, the linear kernel function with the fastest training speed is chosen to estimate the daily volume for the next month's sorting operations.

3. The Support vector machine model for cargo forecasting after changing sorting transportation routes

3.1. The application of graph theory and network analysis to construct features

The large data prediction model for the user's electricity consumption is implemented in the Clementine software. This paper utilizes graph theory and network analysis to understand the connectivity between sorting centers and the flow of cargo volume. The principle of the network graph visualization algorithm is to display nodes and edges from the data in graphical form, facilitating a more profound understanding and analysis of inter-data relationships [7-8]. Its primary principles are as follows.

(1) Nodes represent entities or objects within the data. Specifically in this paper, they refer to sorting centers included within logistics transportation routes.

(2) Edges denote the relationships or connections between nodes. Specifically in this paper, they represent the cargo volume originating from and arriving at sorting centers.

(3) Layout algorithms determine the position of nodes within the graph, aiming to render the graph more comprehensible and analyzable. This paper employs the shell layout.

(4) Visualization presents nodes and edges using a graphical interface.

Some sorting centers are not involved in interactive sorting transportation is shown in fig.3. For data manipulation, 'Origin Sorting Center' and 'Destination Sorting Center' were filtered using Excel software. The results are displayed in Table.1. It is speculated that these sorting centers may not serve as parcel transfer stations, and thus to accurately grasp the scale of each sorting center and enhance the model's predictive performance, the model training excluded these missing data.

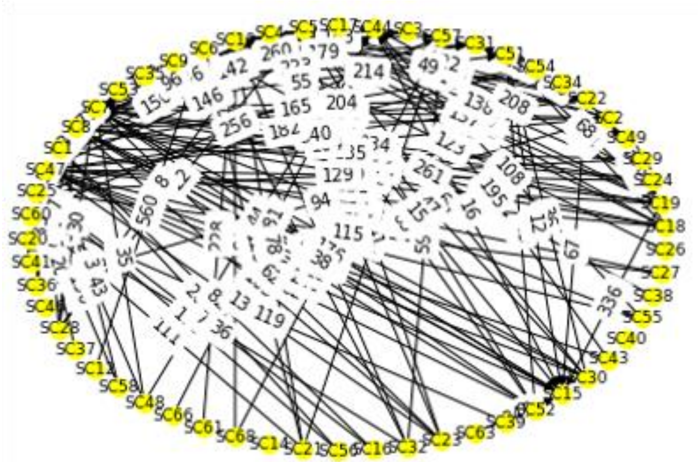

Fig 3. The directed network graph based on shell layout.

Table 1. Selection of missing centers in interactive circuit.

Missing centers before routes change	Missing centers after routes change
SC20	SC20
SC46	SC46
SC14	SC61
SC40	SC14
SC49	SC39
SC17	SC40
	SC49
	SC17

This paper constructs specific features representing the scale of a sorting center, receiving cargo volumes from other centers and dispatching volumes to others. The feature vector includes the average historical values of sent and received cargo volumes for all sorting centers within the interactive sorting network.

3.2. Analysis of experimental results

This paper employs the Particle Swarm Optimization (PSO) for parameter tuning, utilizing optimal c and γ parameters to construct the SVR model. The fundamental concept of PSO is inspired by the food-searching behavior of bird flocks. The PSO algorithm treats solutions to optimization problems as massless particles flying in a D-dimensional space. Each particle possesses a fitness value determined by an optimized objective function, and its velocity dictates the direction and distance of flight. Particles update their position and velocity based on both their own previous optimal positions and the collective best position achieved by the swarm, thus converging towards the global optimum.

Particles update their velocity and position according to the equations below, where $v_{id}(t)$ is the 'flight' velocity of particle i in dimension d, and P_{id} is the individual best position coordinate for particle i in dimension d [9].

$$v_{id}(t+1) = \omega v_{id}(t) + c_1 r_1\big(P_{id}(t) - x_{id}(t)\big) + c_2 r_2\big(P_{id}(t) - x_{id}(t)\big) \tag{5}$$

$$x_{id}(t+1) = x_{id}(t) + v_{id}(t+1) \tag{6}$$

In this study, we define a swarm of 50 particles using a PSO algorithm with a maximum iteration count of 20 to solve for the optimal parameters. The range of parameters considered is c = 0.1, 1.1...100 and γ = 0.01, 0.11...10. The optimal c and γ parameters obtained and their corresponding Mean Absolute Error (MAE) are presented in Table.2. Using the optimal parameters, an SVR model is created, incorporating the average feature constructed in step two as one of the input features for SVR model training. The prediction accuracy evaluation results are shown in Table.3.

Table.2. Optimal parameters and metrics for PSO solving

Optimal parameters		Measuring indicators
c	γ	MAE
100	0.01	8817.19

Table 3. The evaluation of SVR model accuracy.

/	Linear Kernel
MAE	5608.2
R^2	0.5703

4. The application of Integer Planning to schedule crews

This study defines decision variables as follows, the number of permanent workers R_{ijk} attending each shift at each sorting center every day, where i represents the sorting center, j represents the date, and k represents the shift; the number of temporary workers T_{ijk} attending each shift at each sorting center every day, where i represents the sorting center, j represents the date, and k represents the shift.

The objective of this paper is to minimize total personnel costs [10] while meeting the daily cargo volume requirements of each sorting center. Since the employment cost of permanent workers is lower than that of temporary workers, this paper sets different unit costs $Cost_R$ and $Cost_T$ ($Cost_T > Cost_R$) for permanent and temporary workers, respectively. The objective function is as follows

$$min\, z = \sum_{i=1}^{57}\sum_{j=1}^{30}\sum_{k=1}^{6}\left(Cost_R * R_{ijk} + Cost_T * T_{ijk}\right) \tag{7}$$

Ensure the daily cargo volume requirements for each sorting center and every shift are met (each shift lasts 8 hours).

$$\left(R_{ijk} + T_{ijk}\right) * E_i \geq Q_{ijk} \tag{8}$$

$$E_i = \frac{R_{ijk}}{R_{ijk} + C_{ijk}} * 8 * E_R + \frac{C_{ijk}}{R_{ijk} + C_{ijk}} * 8 * E_T \tag{9}$$

E_R represents the hourly efficiency of permanent workers, while E_T represents that of temporary workers. Hourly efficiency is defined as the amount of cargo sorted per person per hour. Substituting $E_R \leq 25$, $E_T \leq 20$ yields as follows

$$8 * \left(\mathrm{R_{ijk}} * \mathrm{E_R} + \mathrm{T_{ijk}} * \mathrm{E_T}\right) \geq \mathrm{Q_{ijk}} \tag{10}$$

Herein, Q_{ijk} signifies the predicted cargo volume during each shift over 30 days following a change in the transportation route.

Each sorting center has 60 permanent workers (the total number of permanent workers attending each shift cannot exceed 60 people).

$$\sum_{k=1}^{6} R_{ijk} \leq 60 \tag{11}$$

Ensure the actual hourly efficiency is as balanced as possible daily (the difference in actual hourly efficiency between adjacent days must not exceed a specific value).

$$\left|\frac{\sum_{k=1}^{6}(\mathrm{R_{ijk}} + \mathrm{T_{ijk}}) * \mathrm{E_i}}{8} - \frac{\sum_{k=1}^{6}\left(\mathrm{R_{i(j+1)k}} + \mathrm{T_{i(j+1)k}}\right) * \mathrm{E_i}}{8}\right| \leq \varepsilon_1 \tag{12}$$

Herein, ε_1 is the allowed maximum difference. Substituting E_i yields as follows

$$\left|\sum_{k=1}^{6}\left(R_{ijk} * E_R + T_{ijk} * E_T\right) - \sum_{k=1}^{6}\left(R_{i(j+1)k} * E_R + T_{i(j+1)k} * E_T\right)\right| \le \varepsilon_1 \tag{13}$$

The arranged staff should be as few as possible (the total number of people arranged for each shift over 30 days is less than a specific value).

$$\sum_{j=1}^{30}\sum_{k=1}^{6}\left(R_{ijk} + T_{ijk}\right) \le \varepsilon_2 \tag{14}$$

Herein, ε_2 is the allowed maximum difference.

The decision variables cannot be negative.

$$R_{ijk}, T_{ijk} \ge 0 \tag{15}$$

The results of this article, which consider the change in transportation routes, predict the hourly cargo volume for 57 sorting centers over 30 days using Python to create a new label ‘TimeShift’. Based on the 'Hour' column of the prediction results, values are assigned to ‘TimeShift’, resulting in Table.4. With equations 10-16 as constraints, this article establishes an integer programming model using Matlab's intlinprog function and solves it.

Table 4. Forecasted Cargo Volumes with TimeShift (Displaying the Top Five Data).

Center	Date	Hour	TimeShift
SC1	2023/12/1	00:00-08:00	3679
SC1	2023/12/1	05:00-13:00	2250
SC1	2023/12/1	08:00-16:00	1422
SC1	2023/12/1	12:00-20:00	1976
SC1	2023/12/1	16:00-24:00	1545

5. Conclusions

This study performs exploratory analysis on the hourly daily cargo volume data of sorting centers, discerning trends related to e-commerce platforms' consumer promotion measures. Leveraging the Support Vector Regression (SVR) model, it accommodates data noise through slack variables, projecting upcoming month's cargo volume. The Particle Swarm Optimization (PSO) algorithm is employed for global search to identify optimal model parameters, enhancing prediction accuracy and generalization. Integer Programming is applied to address multiple constraints in real-world staff attendance planning. The experimental results show that the integration of PSO with the SVR model significantly improves cargo volume prediction accuracy, while Integer Programming effectively estimates the arrangements for staff attendance, culminating in reduced employment expenditures within logistics network sorting centers. In this paper, there is potential for the model's application across various logistics scenarios, aiming to achieve comprehensive optimization in larger, more complex networks.

References

[1] Li Hongjuan, Lu Tianzhe, Zhu Hancan. National Freight Volume Forecast and Freight Structure Analysis [J]. Statistics and Decision, 2022(18): 5.

[2] Huang Juan, Liu Shuangxian. Forecast of National Freight Volume and Analysis of Freight Structure [J]. Science and Technology & Innovation, 2021(16): 3.

[3] Huang Mengting, Miao Bin. Research on the Complex Airport Ground Staff Scheduling Based on Set Partitioning Model [J]. China Collective Economy, 2023(12): 110-113.

[4] Si Shoukui, Sun Zhaoliang. Mathematical Modeling Algorithms and Applications [M]. 3rd Edition. Beijing: National Defense Industry Press, 2022.

[5] Li Hang. Statistical Learning Methods [M]. Beijing: Tsinghua University Press, 2019.

[6] Song Jindong, Zhu Jingbao, Liu Yanqiong, Sun Wentai, Li Shuilong, Zeng Kuiyuan, Wang Yunlong, Yao Kunpeng, Li Shanyou. High-Speed Railway Earthquake Early Warning Method Based on Support Vector Machine Prediction Model [J]. China Railway Science, 2022, 43(5): 177-187.

[7] Li Qiuping, Jia Jingnan, Zhang Yunfei, Luan Xuechun, Ma Ruofei. Analysis of Inter-City Freight Flow Characteristics Based on Freight Trajectory Data [J]. Surveying and Mapping Geography, 2022, 47(5): 84-88.

[8] Zhou Wentai, Yang Ruda, Hou Xinshuo. Analysis of the World Trade in Services Network - Based on Binary/Weighted Perspective and the Quadratic Assignment Problem (QAP) Method [J]. International Trade Issues, 2020(11): 18.

[9] Xie Xue, Chen Junfeng, Zheng Xiuqing, et al. A Freeze-Thaw Soil Evaporation Forecast Model Based on Principal Component Analysis and Particle Swarm Optimization of Support Vector Machine [J]. Water-Saving Irrigation, 2020(1): 6.

[10] Li Bicheng. Research on the Optimization of Logistics Distribution Cost Based on Integer Programming Method [J]. China Logistics and Purchasing, 2022(19): 107-108.