Volume forecast of e-commerce sorting center based on adaptive hybrid ARIMA-LSTM and fuzzy logic

*Canpeng Wu, Shuhui Li**

Guangdong Polytechnic Normal University, Guangdong, China

ABSTRACT

With the rapid development of e-commerce, the logistics sorting center plays an important role. It not only realizes the rapid circulation of goods, but also is the key to improve customer satisfaction and supply chain efficiency. This research is devoted to constructing an accurate forecasting system, which can improve the efficiency of operation, promote the accurate planning of resources and save the cost by forecasting the daily and real-time cargo volume of the Sorting Center, and ensure the timely delivery of goods. This not only enhances the consumer experience, but also optimizes inventory management to prevent inventory shortages, while strengthening risk management and strategic decision-making to promote sustainable development and avoid waste of resources, and gain an edge in the fierce market competition.

In order to improve the accuracy of the prediction model, LSTM long-term and short-term memory network is introduced into the existing Arima model. Arima model is good at processing time series data with obvious trend and seasonality, while LSTM is good at capturing complex nonlinear relationships and long-term dependencies. By combining these two methods, we create an adaptive hybrid ARIMA-LSTM prediction model.

The results show that compared with Arima model alone, the adaptive hybrid ARIMA-LSTM model has a significant improvement in the prediction effect, and the prediction accuracy has been greatly enhanced. By introducing LSTM, the model can better understand and deal with the non-linear characteristics and long-term dependencies in time series data, thus making the prediction more accurate, this is of great value in guiding the operational strategy and resource planning of the Sorting Center.

By comparing the regression fitting between D1 and D2 to D5, it is found that the adaptive hybrid ARIMA-LSTM model is more advantageous than the single Arima model in the prediction of cargo volume. Thus, the model can be used to predict the future 24 hours of any day at every moment of the volume, to provide more detailed data guidance. However, because of the complexity and uncertainty of such predictions, we also need to further modify them using fuzzy logic to address possible inaccuracies or ambiguities.

Keywords: ARIMA Models; LSTM Long Short-Term Memory Network; Visualized Analysis; Fuzzy Logic; Ftting Analysis

1 RESEARCH BACKGROUND

With the rapid development of e-commerce, people are more and more used to online

shopping, which makes e-commerce platform of the rapid increase in the flow of goods, to the sorting center a huge challenge. In order to meet the increasing demand of transportation and storage, the sorting center must improve its sorting efficiency, optimize the utilization of resources, and ensure the accurate and timely delivery of goods. An efficient e-commerce sorting center not only helps the goods to arrive at the destination more quickly and accurately, but also can arrange the staff's shifts reasonably according to the quantity of goods, and improve the economic benefits, by maximizing the use of time and costs to reduce the rate of waste of resources and promote sustainable development [1].

In the Sorting Center, the volume forecast is a key link. Accurate volume forecasting can improve the management efficiency of enterprises and reduce unnecessary costs, planning well in advance, so that goods can not be delivered to the hands of consumers due to the backlog of the phenomenon, better forecasting of supply chain changes. In the supply chain, there are many links to maintain this one industry chain, each link changes will affect the overall operation of enterprises. Therefore, do a good job of volume forecast, prepare well in advance, can effectively use resources to complete more work [2]. It is one of the effective ways for e-commerce enterprises to promote their core competitiveness to meet the challenge through goods forecast.

E-commerce sorting center as the main core step in the logistics process, consumers have higher requirements for it. In the sorting process, goods because of their own materials, shelf life and other conditions, the need for sorting center with different ways to classify and preserve. This requires e-commerce sorting center to enhance flexibility, classification, constantly adjust their sorting methods and distribution methods to meet the changing shopping needs of consumers.

In this intelligent era, e-commerce sorting center in the gradual introduction of artificial intelligence. Through artificial intelligence, the sorting center can not only improve the ability of data processing and analysis, improve the ability of forecasting goods, but also realize the maximum use of resources, and enhance the service quality of the Sorting Center, enhance the competitiveness of enterprises [3]. Artificial Intelligence (AI) can also analyze historical transaction data. Through the adaptive hybrid model of AI, the data can be preprocessed in different scenarios to obtain meaningful information. Combining different information to get the valuable rules of goods sorting, improve the management efficiency of the sorting center.

2 SOURCE OF DATA SET

E-commerce logistics network consists of many links, among which the sorting center is the middle link of the logistics network. The quantity of goods in a sorting center usually refers to the total quantity of goods handled through the sorting center in a certain period of time. This volume index reflects the volume of logistics handled by the sorting center in the e-commerce logistics network, including all types of parcels and goods. Volume can be measured on a daily, monthly, quarterly, and yearly basis [4].

Therefore, through national data, CEIC, search number, 199IT and other relevant data collection sites for data retrieval and collection, in the process of searching, it is found whether the data exist, whether the data is reliable, outdated, accurate, complete, general and

multi-interpretation, lost the original details and accuracy, whether the data are geographically different, whether the data is representative, and data copyright and privacy issues. Based on the above principles, find two volume data sets.

Data from the 14th MathorCup mathematical applications challenge. The data is the daily and hourly volume of 57 sorting centers for three consecutive months. The source year of the data set is 2023 to meet the timeliness of the data, because the time is close, the amount of data is big, the generalization is strong, the detail is accurate, therefore this data has certain representativeness [5].

The data includes the actual volume at each time and the average volume for each route. The data contains four data sets:

The first is the daily volume of 57 sorting centres with fixed transport routes between the centres for four consecutive months (August, September, October and November 2023) , as shown in volume data set 1.

The second is November 2023, the month in which the 57 sorting centers had a daily hourly volume (24 hours) , as shown in volume data set 2.

Then the two collected volume data set analysis and detection, will find that there are some data sets missing, overlapping problems, for example, the 4-month daily volume of 57 sorting centers in volume data set 1 is missing for several days, such as 2023.9.23 in SC54,2023.11.7 in SC66, and so on. For these missing data, through the aggregate found in the data set in the proportion of less than 0.01%, according to the actual analysis, this is a normal phenomenon, by solving the monthly volume average, fill in the missing data to ensure data integrity of the four data sets for subsequent volume data visualization analysis and related modeling.

2.1 The data is known to analyze the factors that influence the volume of the Sorting Center

E-commerce logistics as the core of today's social supply chain, the sorting center of the fluctuations in the volume of the impact of the growing e-commerce market efficiency. Among them, there are many factors that affect the fluctuation of the sorting center's cargo volume, for example, holidays and promotional activities lead to the volume of goods surge, economic conditions, weather patterns, and even the development of political events and other internal and external factors will affect it.

Through the research of statistics and operations research, we can find that these internal and external factors can be divided into irrelevant variables. According to the data query, the change of transportation routes among the sorting centers is the most influential one, an optimal route can minimize the delivery time of packages and improve the package handling capacity. However, any change in route will have to be carefully considered, as it could lead to significant volume fluctuations [6]. For example, a change of route may bypass some efficient sorting centres, leading to a drop in volume while increasing pressure on others. Therefore, the transportation route is not easy to change, but in the Sorting Center for cargo forecasting, transportation route changes are still common. Therefore, the volume forecast of Sorting Center in e-commerce logistics network can be divided into two situations:

If the transportation route is fixed, other uncertain assumptions do not occur;

If there is a change in the route, other uncertainty assumptions do not occur.

After the data certification, the logistics transportation route is not easy to change, so in this study, do not consider the impact of the transportation route changes [7]. In this paper, a hybrid forecasting model is established to predict the daily volume of the sorting center in time when the transportation route is fixed and there are no other disturbing factors.

2.2 Analysis And Pre-Processing Of Data Visualization

2.2.1Data Visualization

The data of volume data set 1 and volume data set 2 are cleaned, and the whole analysis is carried out to remove all the irrelevant data, and only the continuous data which is beneficial to the later volume prediction model are kept. Based on the data after cleaning, using MATLAB to carry out basic big data statistical analysis, including the average volume, median, distribution and so on, so as to establish the basic understanding of the data, facilitate subsequent judgment and use [8].

For data visualization, first of all, the data of 57 sorting centers is imported into MATLAB editor, and the basic data is analyzed through code, then the output data of the analysis was drawn 57 sorting center volume broken line diagram (the relationship between time and volume) , and then decomposed into 57 sorting center map, see the figure below. Through the two-step operation can be directly seen in the sorting center in this three consecutive months of daily volume changes.

2.2.2 White Noise Test

Based on the visualized graph of the volume-time relationship of the 57 sorting centers, the visual estimation method shows that the volume of the 57 sorting centers is obviously different from the normal volume for several days, these data are defined as abnormal data. Because the visual method is not scientifically rigorous.

Therefore, the use of white noise test data processing, detection of abnormal values, through systematic analysis can be obtained on the day of the 11th double-day shopping volume explosion, the data is a special situation and not normal data, removing these outliers normalizes the volume data [9]. From the analysis, it can be seen that the cargo volume of 57 sorting centers has changed periodically in the past four months, arima free regression moving average model.

2.3 Model Fitting Validation

In order to ensure the accuracy of the forecast result and enhance the persuasiveness of the forecast, it is necessary to carry out a pre-processing test on the Arima free regression moving average model before it is formally used to forecast the cargo volume of the Sorting Center, to ensure that the model is really suitable for the research of the daily and even hourly volume forecast of the Sorting Center in today's e-commerce logistics network.

The process is as follows:

Firstly, the cargo volume of the first three months (August, September and October) in the preprocessed cargo data set 1 is combined with the Arima free regression moving average model.

Then the Order of autoregressive analysis P, the Order of Difference Times D, and the order of moving average components q are calculated based on the daily volume of the three months, and Arima (p, D, Q) is obtained, then we can get the volume value of the fourth month, and set the forecast value to S1, and the daily volume of the fourth month after the actual pre-treatment (November with the outliers removed) to S2.

Finally, the regression fitting function of S 1 and S 2 was analyzed, and the regression fitting degree D 1 of S 1 and S 2 was obtained, and d 1≈0.7573 was obtained [10].

Conclusion: Because D > 75%, the regression fitting accuracy is high, although because the model only considers the linear part of the sorting center, but because of its high regression fitting. Therefore, the Arima free-regression moving average model is suitable for the daily volume prediction of these 57 sorting centers.

3 FIXED TRANSPORTATION ROUTES, THE VOLUME FORECAST MODEL

3.1 Establishment Of Adaptive Hybrid Arima-Lstm Prediction Model

3.1.1 Establishment Of Arima Free Regression Moving Average Model

Arima model is a free-regression moving average model, which is one of the time series analysis methods. The basic idea is to regard the data sequence formed by the prediction object over time as a random sequence and describe the sequence approximately with a certain mathematical model, after the model is identified, the future value can be predicted by the past value of the time series.

Among them:

AR denotes autoregressive model, which is used to forecast and analyze with its own data, and its corresponding parameter-autoregressive analysis order P.

Ma denotes the moving average model: when there is an error term, it can eliminate the random fluctuation in the prediction to reduce the error.

Arma represents moving average model: the use of the premise is that the original data meet the requirements of stationarity, when the error of the autoregressive model is accumulated, the moving average method can effectively eliminate the prediction of random fluctuations.

Arima represents the differential autocorrelation moving average model: on the basis of ARMA model, the differential operation is carried out for the data which does not meet the requirement of stationarity.

ACF autocorrelation function compares an ordered sequence of random variables with itself, which reflects the correlation between the values of the same sequence in different time series.

The PACF partial autocorrelation function calculates the strict correlation between two variables, which is the degree of correlation between two variables after removing the interference of the intermediate variable.

Due to many factors affecting the volume of goods in the sorting center, in addition to the changes in the main transportation routes, then there are internal and external factors,

such as the surge in goods due to holidays and promotions, the state of the economy, weather patterns and even the development of political events, lead to our daily time series volume data showing dynamic changes in the situation, that is, non-stationary data. Therefore, the Arima model is used to carry out differential operation on these non-stationary data to eliminate the dynamic change of the sorting center and seasonal changes, so that the cargo volume time series data presents a stable state.

Arima model has three parameters: autoregressive order P, difference order d, moving average component order Q, which is usually expressed by Arima (p, D, Q) . Observations that define study data are satisfied

$$Z^t = \lambda_1 Z_{t-1} + \lambda_2 Z_{t-2} + \lambda_3 Z_{t-3} + \lambda_p Z_{t-p} \quad (1)$$

Where, for the regression parameter, i = 1,2, ... P is the number of hysteresis variables, and VT is a white noise process, then the ZT of linear data is a p-order autoregressive model, expressed as AR (p) .

White Noise VT is represented by a hysteresis operator

$$v_t = \Lambda(L) z_t = (1 - \lambda_1 L - \lambda_2 L2 - \ldots - \lambda_p L^p) zt \quad (2)$$

(L) is an autoregressive operator.

The autoregressive operator is

$$\Lambda(L) = (1 - G_1^{-1} L)(1 - G_2^{-1} L) \ldots (1 - G_P^{-1} L) \quad (3)$$

$G_1^{-1}, G_2^{-1}, \ldots, G_P^{-1}$Is the characteristic root of the autoregressive characteristic equation. When the characteristic equation satisfies (l) = 0, the AR model is stable in P order.

If the observations are satisfied

$$Zt = vt + \theta_1 v_{t-1} + \theta_2 v_{t-2} + \theta_3 v_{t-3} + \ldots + \theta_q v_{t-q} \quad (4)$$

Among them, $\theta_1, \theta_2, \ldots, \theta_q$ is the white noise corresponding to the equation parameter v_{t-q}. The observed z_t is the q-order moving average model, represented as MA (q) .

The autoregressive equation morphs into

$$Zt = \theta(L) v_t = (1 + \theta_1 L + \theta_2 L^2 + \ldots + \theta_q L^q) v_t \quad (5)$$

θ(L) is the moving average operator.

The characteristic equation of the moving average operator is

$$\theta(L) = 1 + \theta_1 L + \theta_2 L^2 + \ldots + \theta_q L^q = 0 \quad (6)$$

The variable for the moving average operator is

$$\theta(L) = (1 - H_1^{-1} L)(1 - H_2^{-1} L) \ldots (1 - H_q^{-1} L) \quad (7)$$

Thus, study data observations $k_1, k_2, \ldots, k_q$ are constants.

When the characteristic equation satisfies the MA model, it is invertible in qorder.

Arma model is composed of AR model and MA model, and the expression is:

$$Zt = \lambda_1 Z_{t-1} + \lambda_2 Z_{t-2} + \ldots + \lambda_p Z_{t-p} + vt + \theta_1 v_{t-1} + \theta_2 v_{t-2} + \ldots + \theta_q v_{t-q} \quad (8)$$

Synthetically, the morphologies of Arma are

$$\Lambda(L)zt=\theta(L)\,vt \tag{9}$$

If the time series does not have stationarity, then the non-stationarity model needs difference processing, then the ARMA model is Arima model

On Logistics Node 5, LSTM recurrent neural network with long and short-term memory shows excellent convergence property. Through the continuous training of the data of this node, we can see that the LSTM network performs well in the ability of prediction and convergence. The loop network has a very special memory unit, which can understand and learn complex patterns in time series, and is especially suitable for dealing with long-term dependencies. Therefore, LSTM network has excellent convergence on logistics node 5, which makes the prediction of future trends more accurate. The MAPE is defined as follows:

For each node, the average performance of the two algorithms against the percentage error of MAPF is shown in the figure below.

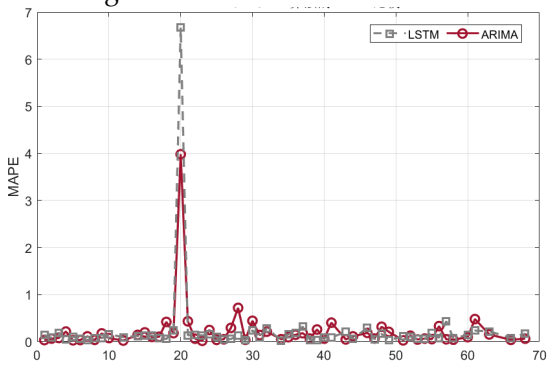

Figure 1: XARIMA algorithm and LSTM model MAPE comparison diagram

There are some differences between Arima and LSTM (long short-term memory) neural network in forecasting effect for different logistics sorting centers. In general, Arima model has advantages in analyzing time series data with obvious trends and seasonality, while LSTM is excellent in dealing with complex nonlinear relationships and long-term dependencies.

Under the application scenario of logistics sorting center, if the data shows periodic characteristics and strong seasonality, such as regular changes by day, week, quarter and year, the Arima model may be a better choice. Arima model can effectively analyze these linear trends and seasonal changes, so as to achieve more accurate forecasting effect.

In contrast, LSTM neural networks are good at dealing with non-linear and dynamic changes, especially for data with complex dynamic characteristics and long-term dependence, LSTM can demonstrate strong modeling and forecasting capabilities. In the logistics sorting center, it may be affected by many non-linear factors, such as sudden demand fluctuation, unexpected situation, etc.

From the evaluation of mean absolute percentage error (MAPE) , Arima model may have better performance for the logistics center with obvious seasonal and periodic characteristics, because it can capture these characteristics better. For the case of complex and dynamic nonlinear model, LSTM model can adapt to the dynamic of data flexibly, so it may have better performance on MAPE.

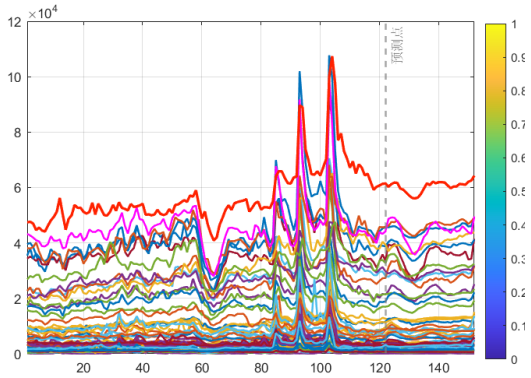

Figure 2: time series data of logistics sorting center

The consistency of the volume distribution and the reasonable trend of the forecast points show that our forecast model can provide more accurate forecast results, has strong reliability for future volume forecasts. It is also worth noting that the distribution of the three peaks in the historical series is different from the rest of the time, so it is necessary to explore and study these peaks.

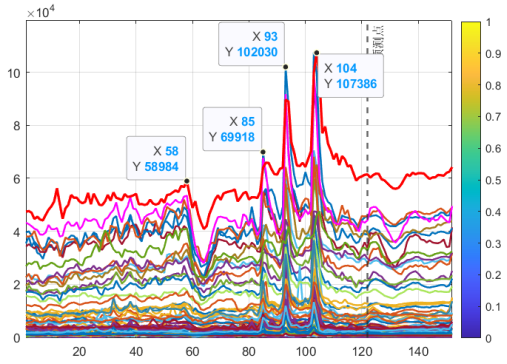

Figure 3: data punctuation chart

From the chart above, we can see that the date of the 85th time point is October 24,2023, and the date of the 104th time point is November 12,2023. Obviously, this period includes the"Double 11" shopping festival. However, during this period, the prediction results of our hybrid prediction model were not affected by these three abnormal peaks, and the model can still correctly identify the overall distribution of samples. This shows that the model has good robustness and accuracy in predicting future volumes.

The results of the adaptive hybrid ARIMA-LSTM prediction model were compared with the actual known cargo data, and the regression fitting degree D2, D3, D4, D5 of each iteration was calculated, finally, the average of these fitting degrees was obtained to be about 0.8677.

Therefore, the adaptive hybrid ARIMA-LSTM model has more advantages than the single Arima model in the volume forecast, and the forecast is more accurate.

REFERENCES

[1] Zhang J, Liu X M, He Y L, & Chen Y S. (2007). Application of ARIMA model in traffic accident prediction (Doctoral dissertation).

[2] Wu J B, Ye L X, & Yuerke. (2007). Application of ARIMA model in predicting the incidence of infectious diseases. Journal of Mathematical Medicine, 20(1), 90-92

[3] Gong G Y. (2008). Application of ARIMA model in GDP forecasting of Shenzhen. Mathematics in Practice and Understanding, 38(4), 53-57.

[4] Chao Zhi, & Sheng Liu. (2019). Comparison of forecasting based on ARIMA model, grey Model and regression model. Journal of Statistics and Decision Making, (23), 38-41.

[5] Gu Jianwei, Zhou Mei, ** Zhi Tao, Jia Xiangjun, & Liang Ying. (2019). Oil well production prediction method based on long short-term memory network model of data mining. Special Oil & Gas Reservoirs, 26(2).

[6] Zhou Xueqing, Zhang Zhansong, Zhu Linqi, & Zhang Chaomo. (2021). A new high-precision fluid identification method based on bidirectional long short-term memory network. Journal of the University of Petroleum (Edition of Natural Science), 45(1), 69-76.

[7] Wen Wen, Tao-Tao Zou, Hong-Yan Wang, & Hai Huang. (2020). Traffic accident prediction model based on dual-scale long short-term memory network. Journal of Zhejiang University: Engineering Science Edition, 54(8), 1613-1619.

[8] Zhang Yufan, Ai Qian, Lin Lin, Yuan Shuai, & ** Zhao Yu. (2019). Region-level ultra-short-term load forecasting method based on deep long short-term memory network. Power Grid Technology, 43(6), 1884-1891.

[9] Fan Tao, XUE Guo , Ping, Yan Bin, Bao Liang, SONG Jin-qiu,... & ** Zerin. (2022). Real-time Inversion method for Deep learning of Transient electromagnetic Long Short-Term Memory Networks. Chinese Journal of Geophysics, 65(9), 3650-3663.

[10] Fu, Jia-Qi, Wu, Hong-Liang, Zhou, Chang-Long, & Julie. (2022). Short-term Stock Forecasting Based on ARIMA-LSTM hybrid Model. Statistics and Application, 11, 630.