Journal of Electronics and Information Science (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/jeis.2024.090108
ISSN 2371-9524 Vol. 9 Num. 1

Time Series Analysis of Product Demand Forecasting and Inventory Optimization on E-commerce Platforms

Ziyu Liu[1,*,#], Yuxiang Zhao[1,#], Shunjie Yang[2,#], Jinli Ju[1], Lianxiang Yang[3], Ruyi Li[1], Jiarong Zhang[4], Weiquan Lu[1]

[1]School of Information Engineering, Xi'an Mingde Institute of Technology, Xi'an, 710124, China
[2]School of River and Ocean Engineering, Chongqing Jiaotong University, Chongqing, 400074, China
[3]School of Management, Xiamen University Tan Kah Kee College, Zhangzhou, 363105, China
[4]School of Intelligent Manufacturing and Control Technology, Xi'an Mingde Institute of Technology, Xi'an, 710124, China
[#]The above authors contributed the equally.
[*]Corresponding author: lzy2398642388@gmail.com

Keywords: ARIMA Time Series Model, K-means Clustering, Cosine Similarity

Abstract: With the continuous advancement of reform and opening up, China's economy has welcomed rapid development, and e-commerce platforms have sprung up like bamboo shoots after rain. The purpose of this study is to use time series models to forecast demand and optimize inventory for thousands of merchants, goods, and supporting warehouses on the e-commerce platform. First, an ARIMA time series model is established for the shipment of old products over time, and through continuous iteration, the optimal parameters of the time series model are obtained for predicting the old products. Then, using K-means clustering, the final prediction results are categorized. Later, new products replace the old ones, and after extracting the feature values of both new and old products to conduct cosine similarity analysis, adjustments are made to the new prediction model to obtain the final forecast values.

1. Introduction

With the rapid development of China's national economy and the continuous advancement of information technology in the national economic and social development[1], the Chinese e-commerce retail industry has achieved impressive results[2]. Therefore, this research aims to respond to the national call by forecasting products and optimizing inventory for e-commerce platforms[3][4].

This research divides the data into two sets: new and old products. It begins by identifying the autoregressive order, the number of lags in the moving average part, and the number of differences required for a stationary time series for all old product data[5]. After repeated experiments, the best model parameters are identified as ARIMA(2,1,1). This model is then used to predict the data for the next 15 days[6][7]. The results are categorized into high-demand and low-demand products using K-means clustering[8].

Finally, the study substitutes new product data into the original data and extracts seven feature

values from both new and old products — including mean, median, standard deviation, maximum value, minimum value, trend characteristics, and autocorrelation characteristics — and performs a cosine similarity analysis[9]. It is found that the introduction of new products does not affect the optimal parameter values of the ARIMA time series model. Therefore, the parameters are kept unchanged, and the future 15 days of product shipment volume is predicted. The Source of Data: http://www.mcm.edu.cn/index_cn.html.

2. Time Series Analysis and Product Demand

2.1 Establishment and Solution of ARIMA Time Series Prediction Model

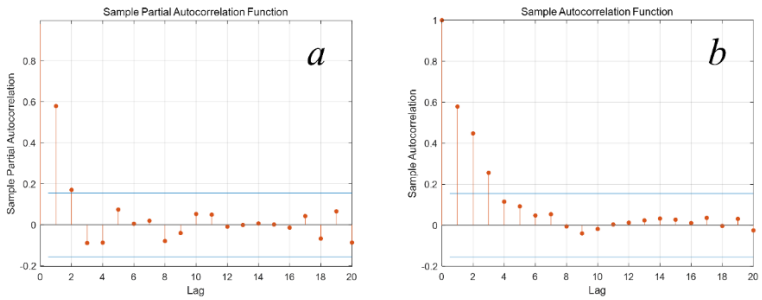
When analyzing the data, it's important to note that the order of the old product data has been randomized. Therefore, the first step is to integrate the data. For example, the first product has the following identifiers: seller_no: 19, product_no: 448, warehouse_no: 30.

In the data preprocessing results, variables such as demand, product category, seller category, inventory category, seller level, warehouse category, and warehouse area have been organized into daily interval data. Based on the *ARIMA*(*p*,*d*,*q*) model, an analysis of the relationship between sales volume and variables like product category is conducted. This analysis is used to predict the demand for products from various sellers in different warehouses for the period from May 16, 2023, to May 30, 2023, denoted as demand x_t. The specific formula is as follows:

$$x_t = \Phi_1 w_{t-1} + \Phi_2 w_{t-2} + \cdots + \Phi_p w_{t-p} + \mu_t + \theta_1 \mu_{t-1} + \theta_2 \mu_{t-2} + \cdots + \theta_q \mu_{t-q} \tag{1}$$

Herein, Φ represents the autoregressive coefficients, w represents the result after d-th order differencing at time t, μ_{t-q} represents the white noise term at different time points, and θ represents the coefficients of the moving average equation.

In the time series, the value of d can be determined through the ADF (Augmented Dickey-Fuller) test to judge whether the time series is stationary; the value of p can be determined by examining the Partial Autocorrelation Function (PACF) plot; the value of q (the number of lags in the moving average part) can be determined by examining the Autocorrelation Function (ACF) plot. The ADF test can determine that the value of d is 1. The PACF and ACF test graphs are shown in Figure 1 below.



a: The Partial Autocorrelation Function Test; b: The Autocorrelation Function Test

Figure 1: Time Series Diagnostic Plot.

Through the ADF test, it can be determined that the value of d is 1; from Figure 2, the Partial Autocorrelation Function (PACF) indicates the value of p is 2 or 3; and from Figure 1, the Autocorrelation Function (ACF) indicates the value of q is 1 or 2; hence, there are four possible combinations. After comparing multiple results, the *ARIMA*(2,1,1) model is chosen for its more stable forecasting process and results, which means the autoregressive (AR) part has 2 lags, differencing is 1, and the moving average (MA) part has 1 lag in the ARIMA parameters. The accuracy of the model

has been validated. Therefore, the final parameters for the time series are determined to be *ARIMA*(2,1,1). Under the time series model that has passed the accuracy test, the final demand forecast results are obtained, with some shown in Table 1 below:

Table 1: Partial Display of Time Series Forecasting Results.

seller_no	product_no	warehouse_no	date	forecast_qty
2	65	1	167	5.296621
2	65	1	168	5.449341
2	65	1	169	5.396025
2	65	1	170	5.403612
2	65	1	171	5.421008

2.2 ARIMA Time Series Model Forecasting Performance Evaluation

A commonly used indicator for assessing the accuracy of forecasts is 1-wMAPE:

$$1-wampe=1-\frac{\sum\left|y_i-y_i\right|}{\sum y_i} \tag{2}$$

Where y_i is the actual demand of the i th series (the daily quantity of various goods stored by merchants in each warehouse), and y_i is the predicted demand of the i th series.

By adjusting the original time series model to predict the last ten days of data, i.e., the shipment data of various products in various warehouses from May 6, 2023, to May 15, 2023, and comparing it with the original data using this formula, the accuracy rate of this time series prediction model is calculated to be 68.78%. Considering the back-and-forth changes in the time series and the relatively small amount of time data, this prediction accuracy is within an acceptable range from a statistical standpoint. To facilitate warehouse management, clustering algorithms are used here to categorize the time series.

2.3 Using the K-means clustering method to categorize the results of time series predictions

First, extract the feature values from the time series formed by merchants, warehouses, and products. Then, use the trend feature values and autocorrelation feature values as the quantitative data for the K-means clustering method.

① Trend features: Trend features are calculated through linear regression. Perform linear fitting on the time series data to obtain the slope of the trend. Hence, a linear model is established:

$$y(t)=m\cdot t+c \tag{3}$$

Where $y(t)$ is the value of the time series; t is the time point in the time series; m is the rate of change of the time series (slope); c and is the intercept.

② Autocorrelation features: For a given lag k, the autocorrelation coefficient r(k) represents the correlation between the time series y(t) at time t and t+k. The formula for autocorrelation coefficient r(k) is:

$$r\left(k\right)=\frac{\sum_{t=1}^{N-k}\left(y\left(t\right)-\bar{y}\right)\left(y\left(t+k\right)-\bar{y}\right)}{\sum_{t=i}^{N}\left(y\left(t\right)-\bar{y}\right)^2} \tag{4}$$

$y(t)$ is the value of the time series; $\bar{y}$ is the value of the time series; N is the length of the time series; k is the lag.

③ Results of K-means clustering analysis:

There are three metrics, namely silhouette coefficient, DBI (Davies-Bouldin Index), and CH (Calinski-Harabasz Score), to evaluate the effectiveness of K-means clustering analysis. The final indicator parameters obtained from K-means clustering analysis are shown in Table 2:

Table 2: Parameter Values of K-means Clustering Analysis

Silhouette Coefficient	DBI	CH
0.516	0.766	1272.562

According to Table 4, it is evident that the K-means clustering analysis has yielded highly favorable results. This clustering strategy is well-suited for the time series feature vectors in this particular problem. Through repeated testing, we have categorized these time series into two major groups, with Category 1 representing high-demand products and Category 2 representing low-demand products. The results of the clustering analysis indicate that the data can be divided into 2 clusters. The frequency of Category 1 is 797, accounting for 39.97% of the total, while the frequency of Category 2 is 1197, accounting for 60.03% of the total. To showcase the clustering analysis results for 10 samples selected from Business ID 2, as shown in Table 3:

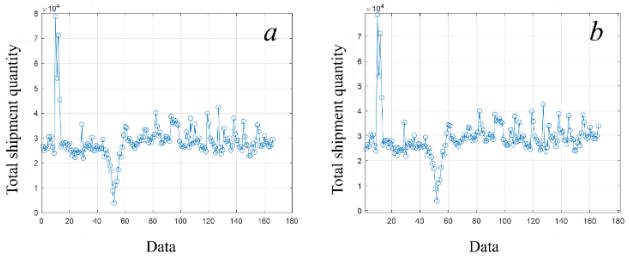
Table 3: Partial Presentation of Clustering Analysis Results

seller_no	product_no	warehouse_no	type	TrendSlope	AutocorrLag1
2	69	10	2	-0.000233487	0.034789212
2	72	1	1	0.062118042	0.424920994
2	73	1	1	-0.090863181	0.723329591
2	74	1	1	-0.075350722	0.666651603
2	74	11	2	-0.004657935	0.281046776

3. Historical Data-Driven Demand Forecasting

3.1 The Establishment of ARIMA Time Series Forecasting Model

To assess the stability of the current dataset, a time series analysis is conducted after introducing new data, aiming to determine the *ARIMA*(*p*,*d*,*q*) parameters. A comparison is made between the ADF test plots of daily shipment data before and after the introduction of new data, as depicted in Figure 2.



a: Before introducing new data into the time series b: After introducing new data into the time series

Figure 2: Time Series Comparison Chart.

From the above Figure 2, it can be observed that the introduction of new data has not had a significant impact on the time series. Therefore, the ARIMA parameters for the time series forecasting model remain as *ARIMA* (2,1,1).

3.2 Comparative Analysis of *ARIMA* Time Series Similarity

Extract seven feature values from both new and old product data: mean, median, standard deviation, maximum value, minimum value, trend characteristics, and autocorrelation characteristics. Use these seven feature values as the coordinate values of the feature vector. Sequences with cosine similarity closest to 1 are considered similar sequences.

The formula for cosine similarity calculation is as follows:

$$cosine\ similarity(A+B)=\frac{A\cdot B}{\|A\|\|B\|} \tag{5}$$

In which, $A\cdot B$ represents the dot product of vectors A B and, $\|A\|$ is the magnitude of vector A, and $\|B\|$ is the magnitude of vector B.

The similar sequences determined through cosine similarity are as follows: The data for the sequences where the cosine similarity between the feature vectors of some old product data and the new product data is closest to 1 are presented in Table 4 below:

Table 4: Partial Presentation of Integrated Data Results

Old product data			New product data			cosine similarity
Seller no	Product no	Warehouse no	Seller no	Product no	Warehouse no	of feature values
11	121	16	1	3	5	0.999607
4	1676	1	1	4	1	0.999659
15	1707	5	1	8	3	0.999712
10	1917	1	1	9	2	0.999960
11	176	19	1	9	5	0.999866

To assign the time data of similar sequences from old products to corresponding products in the new product dataset, and finally, to obtain the forecast values for these dimensions from May 16, 2023, to May 30, 2023, using a time series model with *ARIMA*(2,1,1) parameters, here are some of the results from the time series model predictions after introducing new data, as shown in Table 5:

Table 5: Partial Presentation of Results from the Time Forecasting Model after Introducing New Data

seller_no	product_no	warehouse_no	date	forecast_qty
8	731	5	167	14.4721264
8	731	5	168	14.37446075
8	731	5	169	14.68300183
8	731	5	170	14.86164175
8	731	5	171	15.01981118

4. Conclusions

When determining the order of regression, the number of lags in the moving average component, and the number of differencing steps for old product data, four different parameter combinations were tested repeatedly. Ultimately, *ARIMA*(2,1,1) was chosen as the parameter for time series forecasting. After evaluating the accuracy of the predictions, it was found to be 68.78%. Subsequently, the predicted results were subjected to K-means clustering, dividing the products into 797 high-demand products and 1197 low-demand products. Finally, cosine similarity analysis was applied to the feature values of both old and new products, and it was found that the cosine similarity of the feature values

was generally above 99%. Therefore, new product data was incorporated into the old product forecasting model to predict the shipment quantities for the next 15 days. Here are some of the results as shown in Table 6.

Table 6: The Partial Results of the Time Series Forecasting Model after Incorporating New Data.

seller_no	product_no	warehouse_no	date	forecast_qty
8	731	5	167	14.4721264
8	731	5	168	14.37446075
8	731	5	169	14.68300183
8	731	5	170	14.86164175

References

[1] Wang Man. Digitalization and Pilot Free Trade Zones Boost the Robust Growth of Cross-Border E-Commerce [N]. China Trade News, 2023-12-21 (001).

[2] Lai Lingling, Peng Lifang. Research on the Evaluation of Entrepreneurship Capability of New-type Vocational Farmers in the Context of Rural Revitalization Strategy [J/OL]. Price Theory and Practice, 1-5 [2023-12-24].

[3] Shi Chongwen. Establishing a Bulk Commodity Load Forecasting Model to Support the Modern Power Supply Service System [J]. Agricultural Electric Power, 2023, (10): 41-43.

[4] You Lintian. Research on Commodity Futures Price Forecasting Based on Fundamental Factors [J]. Small and Medium-sized Enterprise Management and Technology, 2023, (15): 41-43.

[5] Geng Yanfang, Li Chao, Liu Xiaopeng. Constructing Time Series Models for Predictive Analysis of Health Resource Allocation in Foshan Area [J]. China Health Standards Management, 2023, 14(19): 79-82.

[6] Sun Xuebo, Liu Ning, Wang Yuanjie, et al. The application of the ARIMA model of time series in the ground noise monitoring system of coal mines [J]. Coal Engineering, 2023, 55(10): 111-117.

[7] Liu He, Li Yanchun, Du Qinglong, et al. A High-water Period Yield Prediction Method Based on Multivariate Time Series Models [J]. Journal of China University of Petroleum (Edition of Natural Science), 2023, 47(05): 103-114.

[8] Su Zhiming. Recognition of Residential Distribution Patterns Based on Cluster Analysis [J]. Surveying and Spatial Information, 2023, 46(11): 190-192+198.

[9] Qiu Songqiang. Eigenvalues and Gradient Descent Algorithm [J]. Advanced Mathematics Research, 2023, 26(03): 36-39+50.