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

Vegetable demand forecasting model based on ARIMA

Yongkuan Li *

School of Internet Economics and Business, Fujian University of Technology, Fuzhou, China, 350000

* Corresponding author: lyk23333@outlook.com

Abstract. Establishing an accurate vegetable demand forecasting model is crucial for fresh supermarkets to optimize replenishment and maximize profits amidst fluctuating market demand driven by the perishable nature of vegetables and changing consumer purchasing decisions. This study takes a certain fresh supermarket as an example, targeting the strong periodicity and seasonality of the vegetable market demand characteristics, using SPSS for data analysis and ARIMA time series modeling. Through stationary testing and white noise testing, the effectiveness and applicability of the model were verified, and an accurate forecast of the total demand for vegetables in the supermarket for the next week was made. This forecasting model provides a reliable basis for the supermarket to formulate replenishment plans, helping it maximize profits in vegetable supply, cope with market demand fluctuations, ensure timely vegetable supply, and meet the constantly changing purchasing needs of consumers.

Keywords: Fresh supermarket; ARIMA; Demand forecasting; Time series.

1. Introduction

With the continuous rapid development of the Chinese economy, there has been a significant shift in consumer attitudes, evolving from mere sustenance to a desire for quality and healthiness in food choices. Vegetables, rich in vitamins and dietary fiber, have become a mantra instilled by parents since childhood. Over the past two decades, the market landscape has undergone a profound transformation, driven by the rise of the internet and e-commerce, leading to a growing trend of online vegetable purchases. One of the primary reasons for this shift is the perceived quality assurance associated with online vegetable shopping. However, vegetables, being perishable, depreciate in value over time due to their susceptibility to decay, with their residual value plummeting to zero after the sales period ends. Even within the limited selling window, customers' price expectations for vegetables are constantly changing, resulting in dynamic fluctuations in the currency they are willing to pay. Accurate prediction of consumer demand and subsequent inventory replenishment based on such forecasts theoretically ensures that all vegetables are sold at the highest price, maximizing profits.

Previous studies have focused on demand forecasting and replenishment strategies in the vegetable market:Yang Jingxuan (2023) utilized a hybrid model combining ARIMA, LightGBM, and NGBoost to predict vegetable sales volume, guiding supermarket replenishment decisions [1]. Srikanth Sankaran (2024) successfully forecasted the daily demand for fresh vegetable products (onions) in the Mumbai wholesale market using the seasonal autoregressive integrated moving average (SARIMA) model [2]. Liu Haichao's (2022) research provided valuable insights into understanding the dynamic changes in the vegetable market and optimizing supermarket replenishment strategies [3]. Rahul Priyadarshi (2019) compared Box-Jenkins autoregressive integrated moving average models with machine learning algorithms (including LSTM and SVR) in vegetable demand forecasting at the retail stage. Results indicated that machine learning algorithms outperformed traditional models, reducing prediction errors, minimizing inventory and product waste, and enhancing daily revenue, thereby offering a more effective demand forecasting tool for the retail industry [4]. Guanzong Mo (2023) conducted in-depth analysis on the relationship between supermarket vegetable sales volume and prices, utilizing methods such as systematic clustering and ARIMA prediction models. This aimed to develop more optimized replenishment and pricing

strategies to maximize profits, given the short shelf life and multifactorial impact on vegetable products sales [5].

In the field of vegetable sales forecasting, the ARIMA model has been a commonly used tool due to its demonstrated effectiveness in multiple studies. However, vegetable sales are influenced by various factors, and considering too many factors may make the model overly complex, challenging to capture the inherent patterns of events, and increase computational difficulty. Therefore, in this study, the ARIMA exponential smoothing method was chosen to predict the total daily sales volume of vegetables in supermarkets over the next week. While this approach does not account for factors such as seasonal effects, weather, or temperature on sales, its advantage lies in simplifying the model, making it easier to implement and understand. The emphasis of this study on identifying whether there are universal patterns in vegetable total sales is crucial for supermarkets to formulate data-driven decision-making strategies. However, the exclusion of other factors may limit the accuracy and applicability of the predictions. Therefore, future research could consider integrating more influencing factors to enhance the accuracy and practicality of forecasts.

2. Methods

2.1. Introduction to ARIMA

ARIMA (Autoregressive Integrated Moving Average) is a classic time series analysis method used for modeling and forecasting time series data. Based on the autocorrelation and partial autocorrelation properties of the time series, the ARIMA model transforms the time series into a stationary sequence and then models it using a combination of autoregressive (AR) and moving average (MA) models. The core idea of the ARIMA model is to difference the time series to convert it from non-stationary to stationary, and then use a combination of autoregressive (AR) and moving average (MA) components to describe the internal autocorrelation and lag properties of the series. The ARIMA model typically consists of three parts: the autoregressive (AR) part (p), the differencing (d) part, and the moving average (MA) part (q), hence denoted as ARIMA (p, d, q) [6].

The modeling process of the ARIMA model includes determining the model's orders (p, d, q), estimating model parameters, and model diagnostics. First, the suitable orders are identified by examining the autocorrelation function (ACF) and partial autocorrelation function (PACF) of the time series. Then, the original sequence is differenced until a stationary sequence is obtained. Next, preliminary estimates of the AR and MA orders are made using the sample autocorrelation and partial autocorrelation functions. Finally, methods such as maximum likelihood estimation are employed to estimate the parameters of the ARIMA model, and the model is tested and diagnosed to ensure its fitting effectiveness and predictive performance.

The ARIMA model exhibits strong flexibility and finds wide applications in fields such as finance, economics, meteorology, and industrial production for time series analysis and forecasting [7]. Its advantages include robust adaptability to various time series patterns, strong interpretability of model parameters, and high accuracy in predicting future data. However, the limitations of the ARIMA model lie in its poor fitting performance for nonlinear and non-stationary sequences, as well as the complexity of parameter estimation and model diagnostics, requiring a certain level of expertise and professional knowledge. Therefore, when applying the ARIMA model, it is essential to choose appropriate models and parameters based on specific circumstances, along with thorough model testing and diagnostics [8].

2.2. Stationary test

The graphical method is a rough way to determine the stationarity of an AR (p) model, while the characteristic root method is an exact method for determining stationarity. The AR (p) model can be abbreviated as

$$\phi(B)x_t = \varepsilon_t \quad (1)$$

Assuming $\lambda_1, \lambda_2, ...\lambda_p$ is a stationary sequence, The characteristic roots of the $\{x_t\}$linear difference equation are substituted into the characteristic equation for any λ_i(iϵ(1, 2, ..., p)), there is λ_i^p-$\phi_1\lambda_i^{p-1} - \phi_2\lambda_i^{p-2}$-⋯-$\phi_p$=0.If all the characteristic roots of the equation are within the unit circle, $|\lambda_i|$<1,i=1,2,⋯, p, then the sequence is a stationary sequence [9].

2.3. Model selects

When establishing a model, there are typically several models passing the stationarity test. At this point, it is necessary to select the relevant models according to the principles shown in Table 1 and use information criteria to determine the optimal order of the model [10].

Table 1. Parameters p, q and determination conditions of model

Model	ACF	PACF
AR(p)	tail off	p-factorial truncation
MA(q)	q-factorial truncation	tail off
ARIMA(p,q)	q order trailing	p order trailing

3. Result

3.1. Data selection and visualization

This study covers the vegetable sales records of a fresh supermarket from January 1, 2023, to June 30, 2023, and organizes them into a panel data. The data meticulously record the sales date, time, vegetable category, item number, item name, sales volume (in kilograms), unit price (price per kilogram, in yuan), and other relevant features for each type of vegetable. These vegetables are divided into six major categories, including cauliflower, leafy vegetables, peppers, eggplants, mushrooms, and aquatic roots and tubers. This data aids supermarket managers in gaining in-depth insights into vegetable sales, devising more accurate procurement plans and promotional strategies, and enhancing sales efficiency and profitability.

Analyzing the daily sales volume of each vegetable can help supermarkets better understand customer buying habits and preferences, thereby optimizing product placement and display methods to improve sales efficiency. Additionally, by analyzing the sales performance of different vegetables, supermarkets can adjust inventory and procurement plans in a timely manner, avoiding overstocking or understocking situations, reducing inventory costs, and minimizing product waste. Moreover, tracking the sales volume of each vegetable facilitates the development of more rational promotional activities and pricing strategies, attracting customers to increase purchase frequency and quantity, thereby boosting overall sales revenue. Importantly, analyzing this data helps supermarkets better predict future market demand, enabling them to make more accurate operational decisions, maintain a competitive edge in the fierce market competition, and consistently provide high-quality vegetable products and services to customers. This is crucial for supermarkets to make informed decisions. The total daily sales volume of the six types of vegetables is depicted in Figure 1.

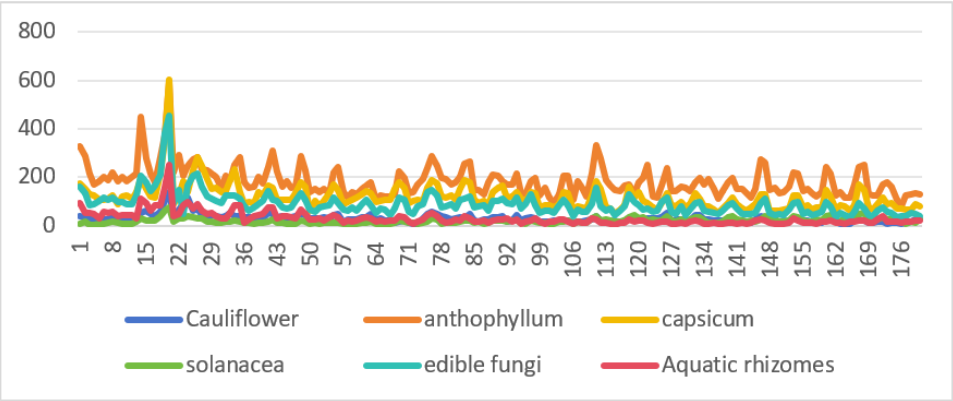

Figure 1. Total daily sales of six types of vegetables

From Figure 1, it is evident that leafy vegetables have the highest total sales volume. Around the 19th day, there is a peak in the sales volume for all six types of vegetables. Subsequently, the fluctuations in sales volume for each type of vegetable are roughly similar. Hence, it can be tentatively inferred that the total sales volume of individual vegetable categories has a relatively minor impact on the overall sales volume of all vegetables.

Next, the proportion of sales for each vegetable relative to the total vegetable sales will be calculated. Analyzing the sales proportion of each vegetable helps the supermarket understand the position and influence of each vegetable in sales, thereby adjusting the product mix and inventory strategy to rationalize the sales structure. By examining the sales proportion of different vegetables, the supermarket can promptly identify popular and slow-moving items, take targeted promotional activities, and adjust procurement plans to increase profitability. Additionally, analyzing sales proportions helps the supermarket understand customer consumption trends and preferences, providing vital insights for optimizing product offerings and services. Overall, calculating the sales proportion of each vegetable serves as a critical basis for supermarket managers to formulate business strategies, enhance operational management, and improve market competitiveness, enabling the market to better meet customer needs and maximize operational efficiency. The area chart displaying the sales proportion of each vegetable is shown in Figure 2.

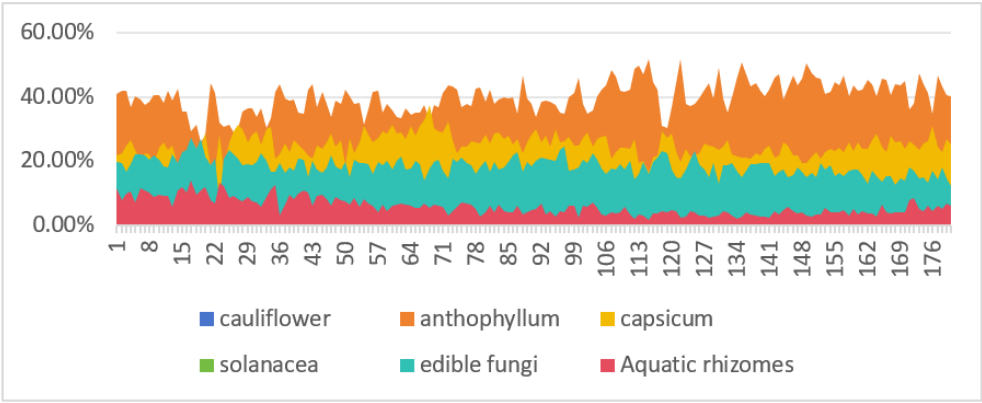

Figure 2. Percentage of sales of each vegetable

From Figure 2, can observe the daily sales proportions of each vegetable. The sales proportions of leafy vegetables, chili peppers, mushrooms, and aquatic vegetables are notably visible, while the proportions for cauliflower and eggplants are too small to be directly observed. Utilizing a box plot helps provide a visual representation of statistical information such as the median, quartiles, outliers, etc. This allows for a quick understanding of the data distribution characteristics and any anomalies present, identifying the central tendency, dispersion, and the presence of outliers, thus enabling more in-depth data analysis and decision-making. The box plot for the total sales volume of all vegetables is depicted in Figure 3.

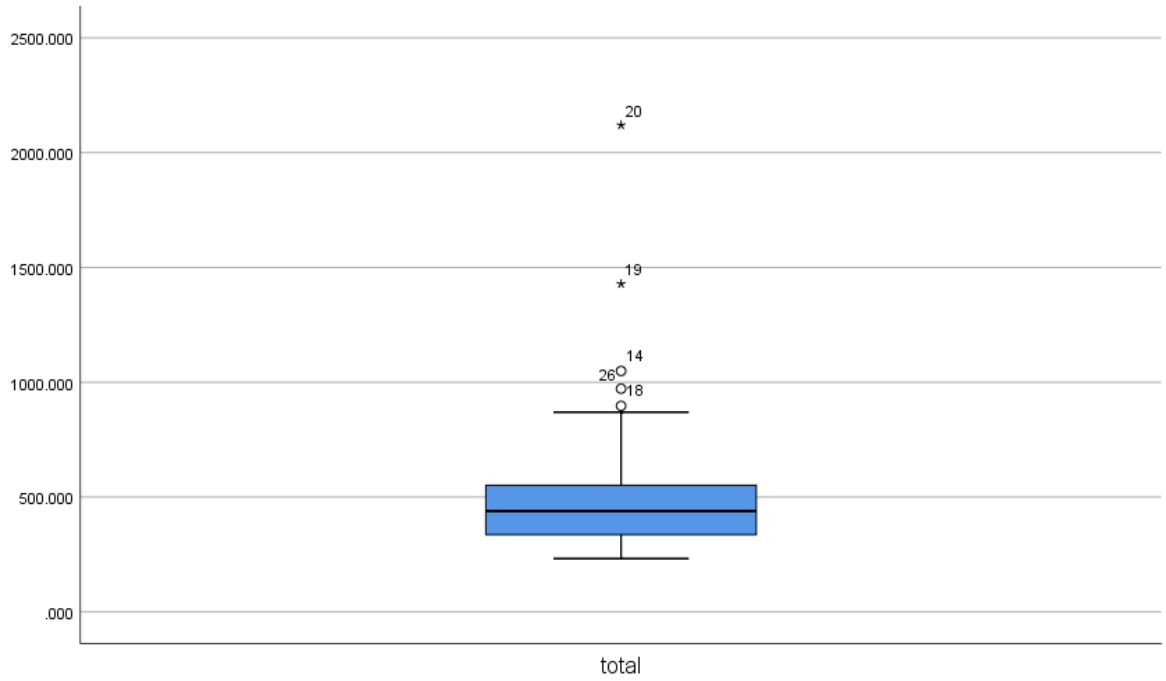

Figure 3. Box plot of total vegetable sales

From Figure 3, observe that the median line is approximately 450, with 5 outliers. These five outliers correspond precisely to the sales peak around the 19th day. Since there was no similar peak observed after the 19th day, can speculate that some event may have occurred on that day leading to a large volume of vegetable purchases by consumers, such as a promotional event held by the supermarket. Given the accuracy of the data collection, there is no need to alter the data.

3.2. Stationary test of sequence

Import the data into SPSS for analysis, obtaining the original sequence plot as shown in Figure 4, the autocorrelation plot as shown in Figure 5, and the partial autocorrelation plot as shown in Figure 6.

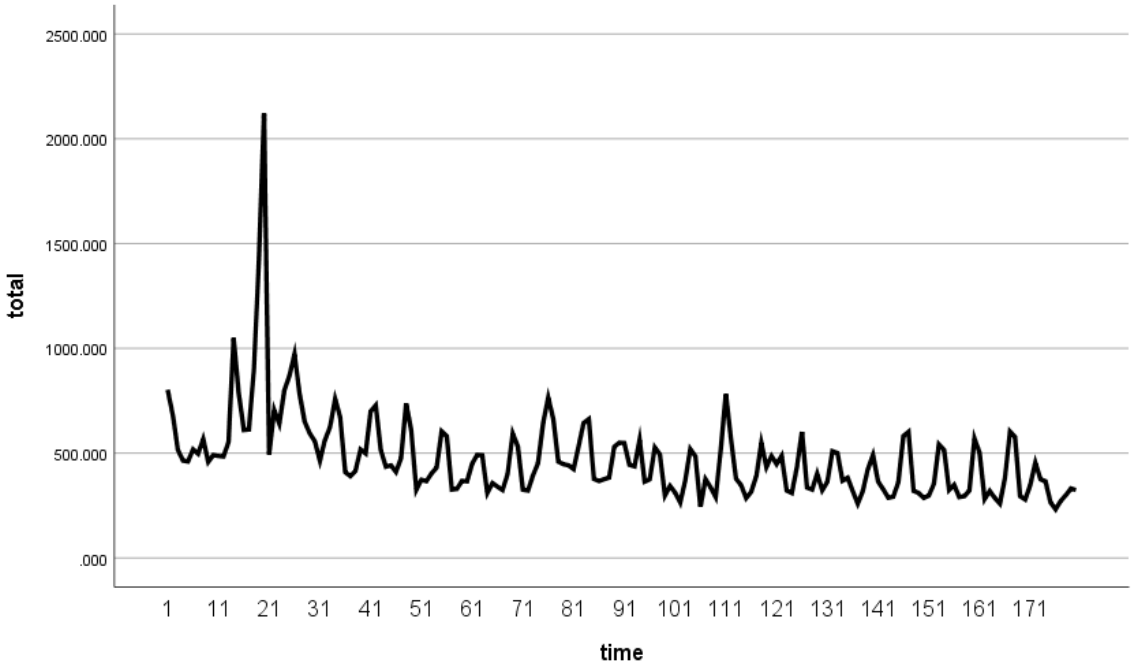

Figure 4. Original time series

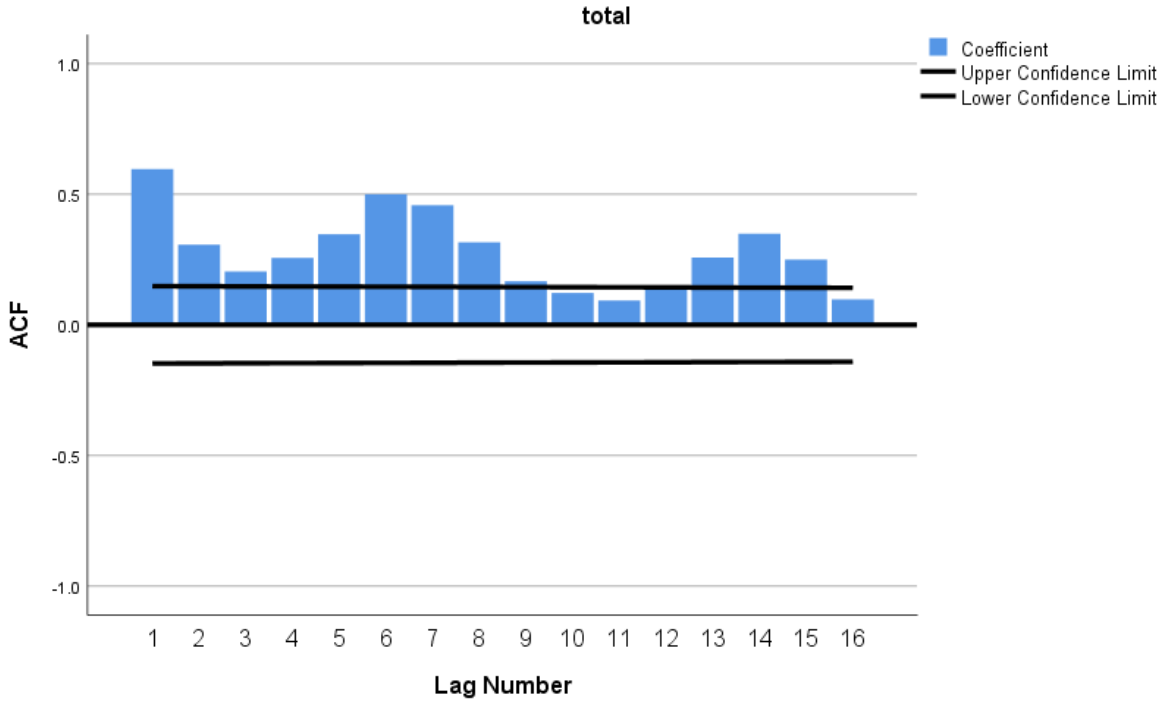

Figure 5. Original time series ACF graph

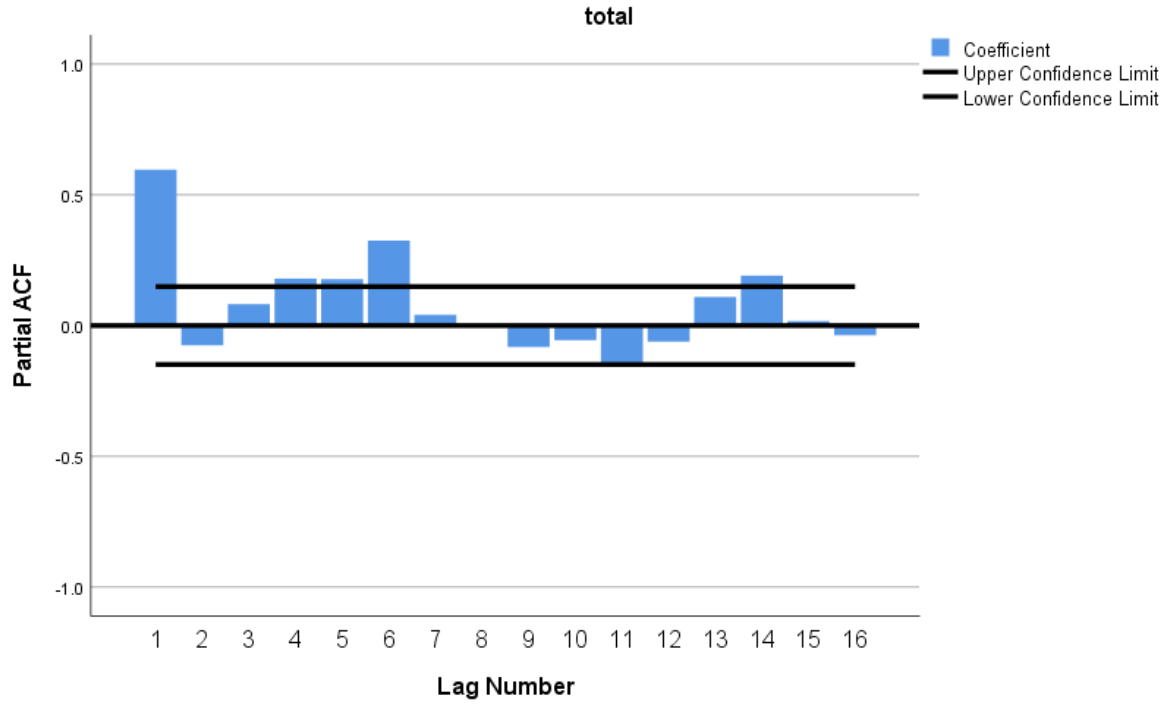

Figure 6. Original time series PACF graph

Based on Figure 5 and Figure 6, it is observed that the ACF fluctuates above the x-axis, while the PACF fluctuates above and below the x-axis. This indicates that the original sequence is not a stationary sequence. When the original sequence is not stationary, it needs to be differenced. By differencing the sequence by n orders, if the differenced sequence becomes stationary, then the differencing order d can be determined.

3.3. Parameter determination

First, the original sequence is differenced by one order. The results are shown in Table 2.

Table 2. ADF test list

variable	Difference order	t	P	AIC	critical value		
					1%	5%	10%
total	0	-1.976	0.297	2133.931	-3.47	-2.879	-2.576
	1	-5.227	0.000***	2124.265	-3.47	-2.879	-2.576
	2	-9.48	0.000***	2135.5	-3.47	-2.879	-2.576

Note: ***, ** and * represent significance levels of 1%, 5% and 10% respectively

From Table 2, it can be observed that when differenced by 0 order, the significance p-value is 0.297, which is not significant at the conventional levels, indicating that the series is non-stationary. When differenced by 1 order, the significance p-value is 0.000***, which is significant at the conventional levels, rejecting the null hypothesis, indicating that the series is stationary.

When differenced by 2 orders, the significance p-value is 0.000***, which is significant at the conventional levels, rejecting the null hypothesis, indicating that the series is stationary. Therefore, the value of d is 1. Next, proceed with the ACF and PACF plot examination to observe their truncation. The results are displayed in Figure 7 and Figure 8.

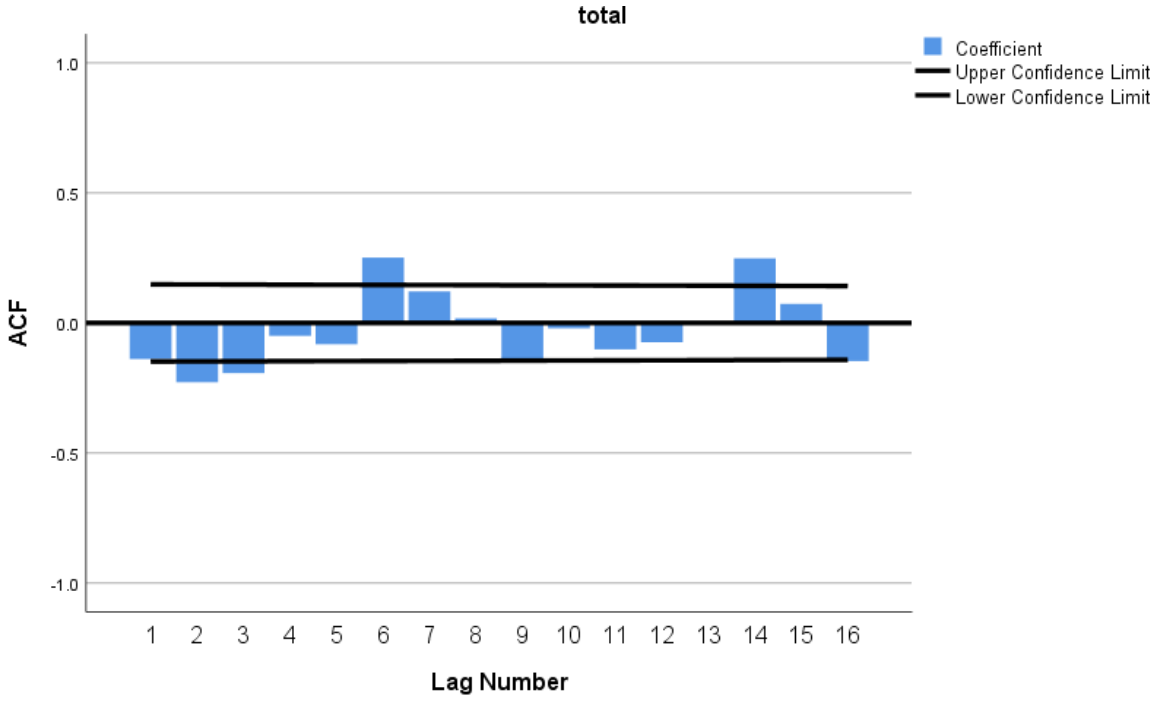

Figure 7. 1st order differential sequence ACF graph

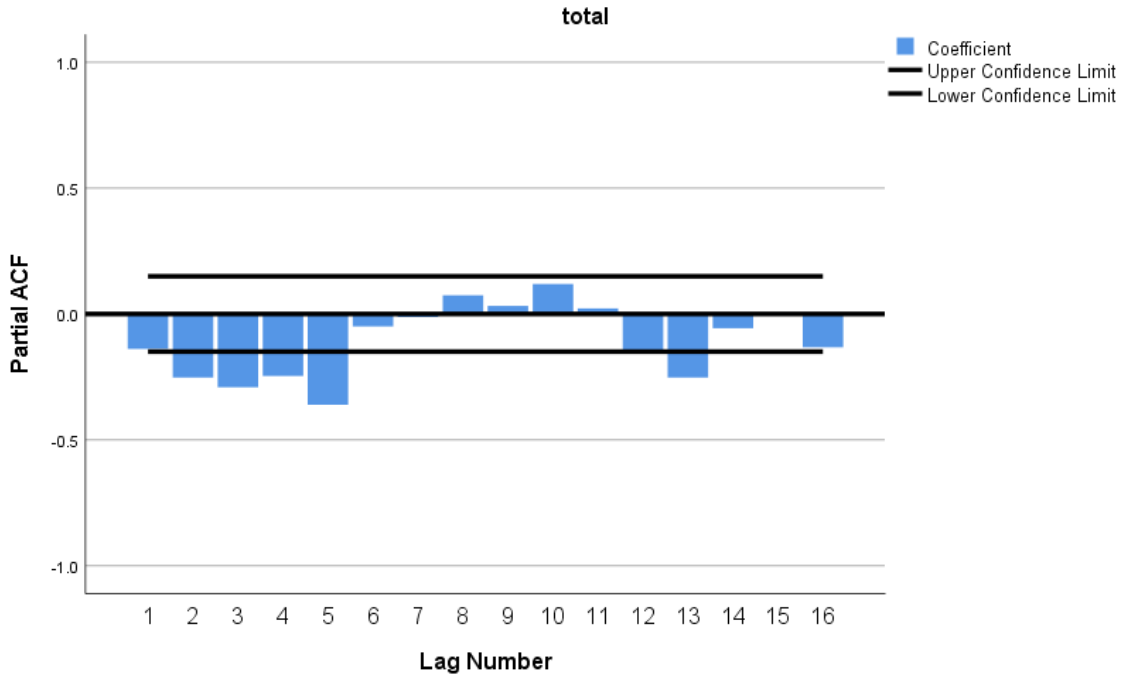

Figure 8. 1st order differential sequence PACF graph

From Figure 7 and Figure 8, it can be observed that after differencing the original series by one order, the ACF exhibits a tail at the 4th lag, and the PACF exhibits a tail at the 6th lag. Most of the ACF and PACF values are within the confidence interval. Therefore, the series after the first-order differencing is stationary. The value of p is determined to be 6, q is determined to be 4, and d is 1.

3.4. Estimation of model

have determined the values of p, d, and q. With the data factors unchanged, proceed to model building in SPSS. The final selected model is ARIMA (6, 1, 4), and the goodness of fit is shown in Figure 9 and Table 3.

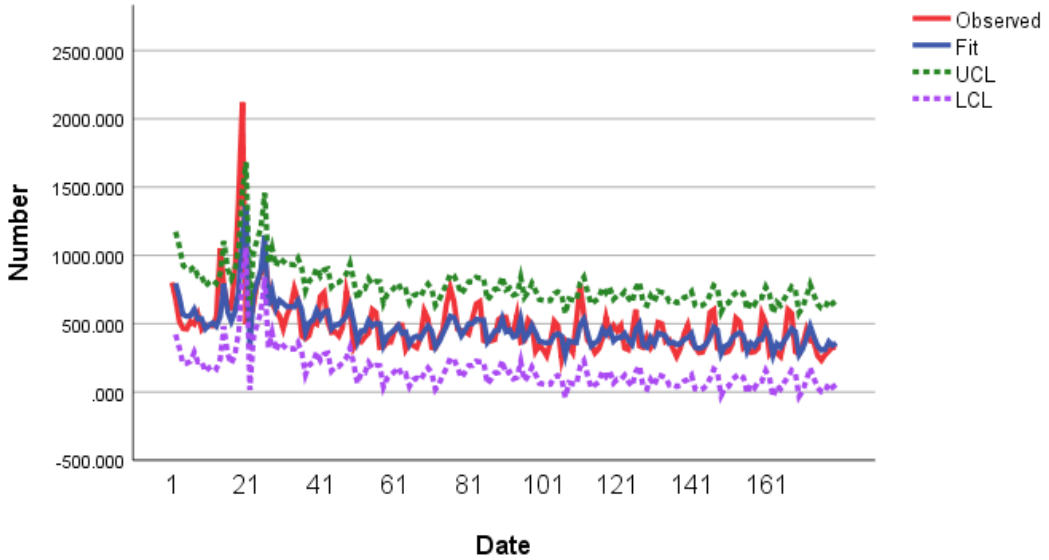

Figure 9. Fitting effect of manual modeling

In Figure 9, can observe that the fitted line closely matches the actual values line, indicating that the model's fit is satisfactory at first glance.

Table 3. ARIMA model check table

item	symbol	value
	Df Residuals	173
sample size	N	180
Q statistic	Q6 (P value)	0.082 (0.775)
	Q12 (P value)	0.92 (0.988)
	Q18 (P value)	13.43 (0.339)
	Q24 (P value)	26.311 (0.093*)
	Q30 (P value)	31.578 (0.138)
Information criterion	AIC	2322.935
	BIC	2345.247
goodness of fit	R^2	0.458

Note: ***, ** and * represent significance levels of 1%, 5% and 10% respectively

From Table 3, based on the Akaike Information Criterion (AIC), the optimal parameters determined automatically for the ARIMA model are (6, 1, 4). The results of the Q statistic analysis based on the variable "Total" indicate that the significance level of Q6 is not significant, suggesting that cannot reject the hypothesis that the residuals of the model are white noise. Additionally, the goodness of fit of the model, represented by R^2 is 0.458, indicating moderate model performance. Overall, the model meets the basic requirements, although there is room for improvement.

3.5. Residual white noise test

After determining the final model, it is essential to conduct a white noise test. The white noise test can also be performed using autocorrelation. The results are depicted in Figure 10.

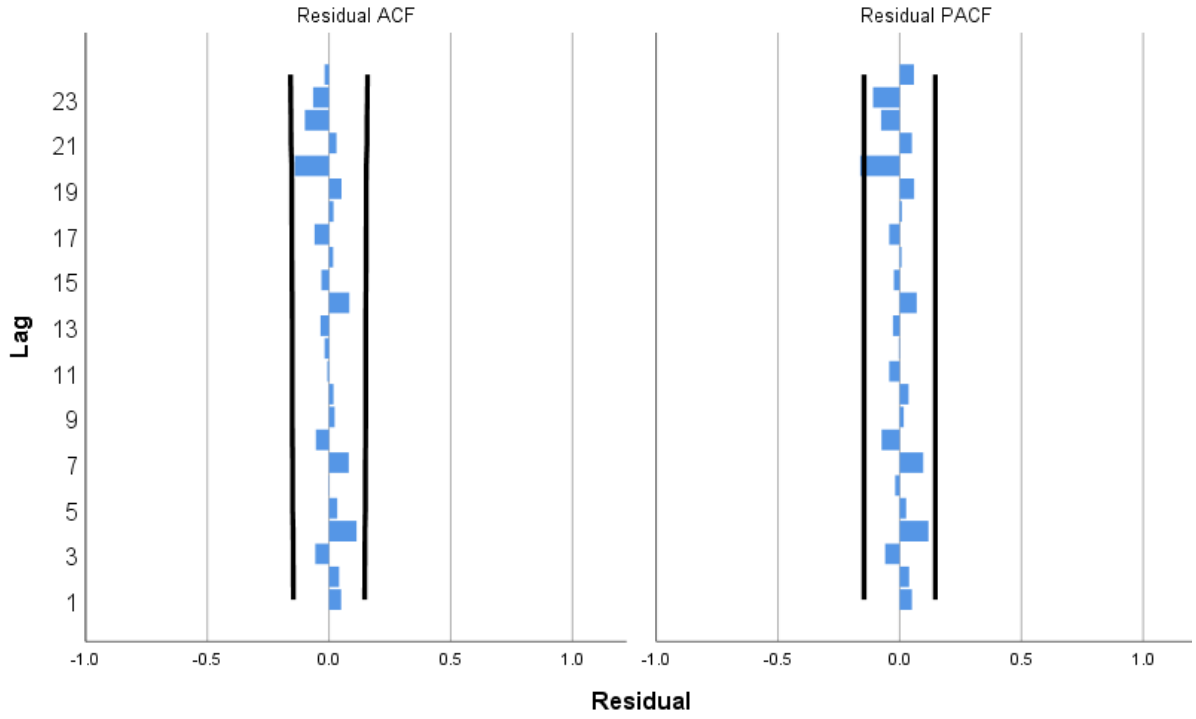

Figure 10. White noise detection diagram

By computing the autocorrelation function (ACF) or partial autocorrelation function (PACF) of a sequence, can assess whether there is significant autocorrelation within the sequence, thereby determining whether it constitutes white noise. If the autocorrelation function of a sequence is not significant at any lag, meaning all correlation coefficients fall within the confidence interval, the

sequence is considered white noise. From Figure 10, it can be observed that all autocorrelation and partial autocorrelation coefficients after first-order differencing fall within the confidence interval. Therefore, it can be concluded that the autocorrelation and partial autocorrelation coefficients are not significant at any lag, indicating that the sequence constitutes white noise. This implies that there is no correlation between observations at different time points in the sequence, and each observation in the sequence is random, independent, and not influenced by past observations. This characteristic aids in forecasting the data.

3.6. Data forecasting

This research will predict the total sales volume of vegetables for the next week. The results are presented in Table 4.

Table 4. Predicted value table

time	forecast the outcome
181	288.16916788812176
182	260.29866180342725
183	266.6861400982257
184	281.6716771764677
185	296.36790706551056
186	297.142357577555
187	283.76543066604745

Based on the data in Table 4, it is observed that the total sales volume of vegetables fluctuates around 300. Through the forecast of total sales volume for the upcoming week, it is noticed that the highest sales volume occurs on Monday, while the lowest is on Wednesday, followed by a gradual increase. Therefore, the supermarket can replenish vegetables based on the sales volume on Monday for the following six days. This strategy can effectively meet customer demand, avoid inventory backlog, and enhance sales efficiency and profitability.

4. Conclusion

The article explores the changes in Chinese consumer attitudes with economic development and the challenges and strategies of the vegetable market in the Internet era. With rapid economic growth, people's demand for vegetables has shifted from simply being satisfied to pursuing better quality and healthiness, reflecting the evolution of consumer attitudes. Meanwhile, the rise of the Internet and e-commerce has changed people's shopping habits, with more and more people choosing to buy vegetables online, mainly because they believe the quality of online-purchased vegetables is guaranteed. However, vegetables, being perishable goods, lose their value over time, and their value becomes zero after the sales period ends, posing challenges to supermarkets, especially in replenishment and pricing. To address this challenge, the article uses ARIMA exponential smoothing to predict the sales volume of vegetables in supermarkets, providing effective replenishment strategies, such as replenishing vegetables for the next six days based on the total sales volume on Monday, to improve sales efficiency and profitability. However, the article does not consider factors such as seasonal effects and weather on vegetable sales, which are areas for further research. In summary, the article provides some insights into understanding the current situation and challenges of the Chinese vegetable market, as well as offering ideas and methods for supermarkets to adopt effective sales strategies in the Internet era.

References

[1] Yang Jingxuan. Based on ARMI-Light... Research on model forecasting of fresh food sales volume [D]. Chengdu: Southwest University of Finance and Economics, 2023.

[2] Srikanth Sankaran. Demand forecasting of fresh vegetable product by seasonal ARIMA model [J]. International Journal of Operations Research, 2014, 20 (3): 315 - 330.

[3] Liu Haichao, Sun Haining. Optimization and simulation of joint decision model for replenishment of warehouse and supermarket based on quantity strategy [C]//The tenth National Youth System Science and Management Science, Xi 'an, Shaanxi, China, 2009: 10 - 17.

[4] PRIYADARSHI R, PANIGRAHI A, ROUTROY S. Demand forecasting at retail stage for selected vegetables: a performance analysis [J/OL]. Journal of Modelling in Management, 2019, 14 (4): 1042 - 1063.

[5] MO G, GUO Y. Vegetable sales forecasting based on nonlinear programming model [J/OL]. Journal of Education, Humanities and Social Sciences, 2024, 25: 190 - 197.

[6] Gao L. Research on Forecasting model of small batch material production demand based on ARIMA [J/OL]. Modern information technology, 2023, 7 (15): 97 - 101.

[7] Jiang Chunhai, YAN Zhenhao, Song Zhiyong. Research on Coal market demand forecast based on ARIMA and GM (1, 1) model [J]. Review of Industrial Economics (Shandong University), 2019, 18 (3): 54 - 86.

[8] Li Baoxin, Xi Qiongqiong. Analysis and prediction of Urban-Rural income gap in Hebei Province based on ARIMA model [J/OL]. Shanxi agricultural economy, 2023, 24: 24 - 28.

[9] Wang Yao. Prediction of China’s Future Population Based on ARIMA Model [J/OL]. Statistics and Application, 2022, 11 (6): 1392 - 1400.

[10] Yan Xiangxiang. The ARIMA model was used to predict the area of park green space [D]. Computer Science, 2023, 47 (S2): 531 - 534.