Research On Time Series Forecasting-Based Pricing Strategy for Supermarkets

Haoke Wang [1], Cheng Zhang [1], Yufei Song [1], Ting Dou [2, *]

[1] College of Field Eengineering, Army Engineering University of PLA, Nanjing, China, 210000

[2] Department of Basics Courses, Army Engineering University of PLA, Nanjing, China, 210000

* Corresponding Author Email: douting163yx@163.com

Abstract. In fresh food supermarkets, vegetables have a short shelf life. It is of great practical significance to price different products based on market research in order to maximize the benefits of supermarkets. The aim of this paper is to investigate and address the pricing decision question of vegetable products in fresh food supermarkets. This article investigates the correlation between cost-based pricing and sales revenue by establishing linear regression models and time series models.In addition, we have also made price predictions for various categories of products in the supermarket. Finally, this article provides pricing decisions for various categories in the first week of July 2023 for the supermarket.

Keywords: Linear Regression, Time Series, Pricing Decision, Supermarkets.

1. Introduction

In fresh food supermarkets, most vegetable products have a short shelf life. Vegetables that cannot be sold on the same day cannot be resold the next day [1]. Therefore, supermarkets make pricing decisions based on the demand and historical sales of various vegetables in order to maximize their profits and minimize waste [2].

In fresh food supermarkets, there is a wide variety of vegetable categories with different origins. Retailers typically engage in buy transactions between 3:00-4:00 am.The supermarket will determine the pricing of recently purchased vegetables by using the cost-plus pricing method [3]. Additionally, vegetables that have incurred transportation losses or have deteriorated in quality will be sold at discounted prices. Reliable market demand analysis is crucial for pricing decisions [4]. From the demand-side analysis [5], there is a certain correlation among the sales volume of vegetable products and time. From the supply-side perspective [6], during the period from April to October, there is a rich variety of vegetable supply, while the selling space in supermarkets is limited. Therefore, determining a reasonable sales combination becomes extremely important.

This study analyzes pricing decisions by establishing linear regression models and time series ARIMA models.The data for this article was sourced from: http://www.mcm.edu.cn/

2. Model Establishment

The linear regression model is a statistical analysis method that utilizes the linear relationship between variables. This study employs a linear regression model as a primary method to examine the correlation between sales volume and cost markup pricing. In addition, an optimization model is constructed with the pricing strategy as the dependent variable, aiming to maximize supermarket profits. Furthermore, an ARIMA model is utilized to forecast the sales of vegetable categories, considering its capability to simulate stationary time series. By incorporating the optimization model, this study aims to predict the necessary conditions for the pricing strategy and facilitate supermarkets in achieving higher profits. The following is the specific process of model establishment.

2.1. Linear Regression Model

The cost-plus pricing method is one of the commonly used pricing methods in the market[3]. It refers to setting the product price to cover the production and sales costs and achieve a reasonable

return. This theory assumes that sellers have the dominant power to determine prices, while buyers can only influence the markup rate. There are two commonly used pricing models for represents the average cost-plus pricing:

The first type: $X = C + Cw$, where X represents price, C represents average cost, w represents the cost markup rate.This model is suitable for pricing products with relatively complete historical cost information.

The second type: $X = C + Tr$, T represents the total capital investment amount, r represents the return on investment rate.This model is primarily used for pricing decisions of new products where there is no historical cost information available. It relies solely on project decision information.

Based on the information obtained through the search, it is understood that vegetable pricing is carried out using the first type of model.

The linear regression equation is a mathematical model used to describe the linear relationship between two or more variables[7]. In this study, it involves one independent variable and one dependent variable, so the linear regression equation can be represented as:

$$y = \beta_0 + \beta_1 x + \varepsilon x \tag{1}$$

Where y is the response variable (also known as the dependent variable), x is the independent variable, β_0, β_1 is the regression coefficient, ε is the error term.

For this study, an analysis is conducted using a regression equation to examine the relationship among the sales volume of six types of vegetables and the cost markup pricing on a monthly basis. A scatter plot of the sales volume against the cost markup pricing is created using Excel functions,

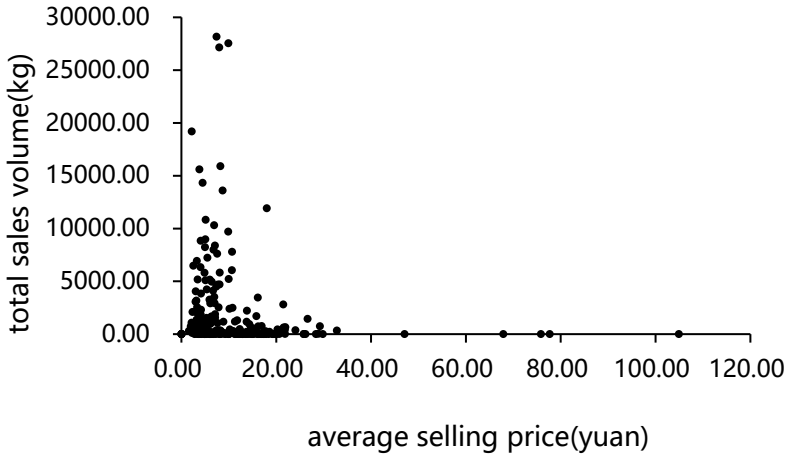


Figure 1 The scatter plot of sales volume against cost markup pricing

Figure 1 shows a scatter plot of sales revenue versus cost-plus pricing,where y represents the total sales volume and x represents the average selling price, through analysis of the image, it is observed that when the sales volume is zero, there is always a cost markup pricing that conforms to market pricing principles. By studying the scatter plot, it is found that when the sales volume is zero, there is always a cost markup pricing of zero, which is in line with market pricing principles. Excluding the case where x=0, Performing regression analysis while protecting the regression function from being affected.

Establishing a linear regression equation:

$$y_i = \beta * x_i + \alpha \tag{2}$$

Substituting the data into the numbered MATLAB program for calculation and analyzing the obtained data such as slope and degrees of freedom. These data analyses can help determine the relationship between the two.

2.2. Time Series Models

Time series analysis is a statistical method used to handle dynamic data[8]. This method studies the statistical patterns followed by random numbers and is used to solve practical problems. It is based on the theory of stochastic processes and mathematical statistics. Time series forecasting models play a crucial role in time series analysis. Time series forecasting methods, as a part of time series analysis, hold significant importance in various fields.

There are two types of known time series models: deterministic and stochastic models. Deterministic models focus on structured forecasting models where the main consideration is given to the trend factors in the time series, while random factors are not taken into account. These models are relatively simpler to compute but may result in decreased accuracy. Deterministic time series models include long-term trend methods, moving average methods, and exponential smoothing methods, among others[9].

Random time series and deterministic time series are different.Random time series takes into account both the influence of temporal factors and the impact of random factors.Compared to deterministic time series models, random time series models have improved accuracy to some extent.The autoregressive integrated moving average (ARIMA) model is an example of a stochastic time series model. In this study, a stochastic time series model is used to analyze vegetable prices.

The ARIMA (Autoregressive Integrated Moving Average) model is a commonly used time series analysis model for predicting future values of time series data. When the data in the time series exhibits an upward or downward trend, it is necessary to transform the data into a stationary time series before building the ARIMA model. The ARIMA model incorporates the concepts of differencing (I), autoregression (AR), and moving average (MA)[10].

The mathematical expression of the ARIMA (p, d,q) model is as follows:

$$Y_t = c + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + ... + \varphi_p Y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - ... - \theta_q \varepsilon_{t-q} \quad (3)$$

2.3. Symbol Explanation

1. Y_t represents the observed values of the time series.

2. c represents the constant term.

3. φ_1, φ_2, ..., φ_p represents the autoregressive coefficient,which signifies the linear relationship between the current observed value and the past p observed values.

4. ε_t represents the error term,which signifies the difference between the observed value and the predicted value of the model.

5. θ_1, θ_2, …, θ_q represents the moving average coefficient,which signifies the linear relationship between the current error term and the past q error terms.

6. p represents the autoregressive order,which signifies considering the past p observed values.

7. d represents the differencing order,which signifies taking the dth difference of a time series to make it stationary.

8. q represents the moving average order,which signifies considering the past q error terms.

3. Experimental Results

3.1. Experimental Results of Linear Regression Model

Linear regression analysis was conducted using MATLAB program. The coefficient results of the linear regression equation are shown in Table 1.

Table 1 The coefficient results of the linear regression equation

Coefficient	Results
Number of observations	251
Error of freedom	249
Root Mean Squared Error	4.13e+03
R-squared	0.0195
Adjusted R-squared	0.0156
F-statistic vs.constant model	4.96
P-value	0.0269

The linear regression analyst is conducted using MATLAB yielded the following results: The slope is -50.512, The error degrees of freedom are 249, with 251 observations, The R-squared value is 0.0195, indicating that the regression model explains1.95% of the variability in the dependent variable. The adjusted R-squared value is 0.0156, The F-statistic (constant model) is 4.96, The p-value is 0.0269.

The variable F is used for variance testing and can be used to test the significance of the entire regression relationship. If $P<0.05$, it indicates that the regression equation is significant and the independent variable has a significant impact on the dependent variable.

In conclusion, it can be inferred that there is a negative correlation between cost markup pricing and sales volume, but the correlation is not significant.

3.2. The experimental results of the time series model.

In the ARIMA(p,0,q) model, experiments were conducted to determine the optimal values of p and q, resulting in the prediction of pricing strategies for vegetables from July 1st to July 7th. Taking into account the periodic nature of time series data, an analysis was conducted on the data around this time period over the past three years. The pricing prediction graph obtained using MATLAB is shown in Figure 2.

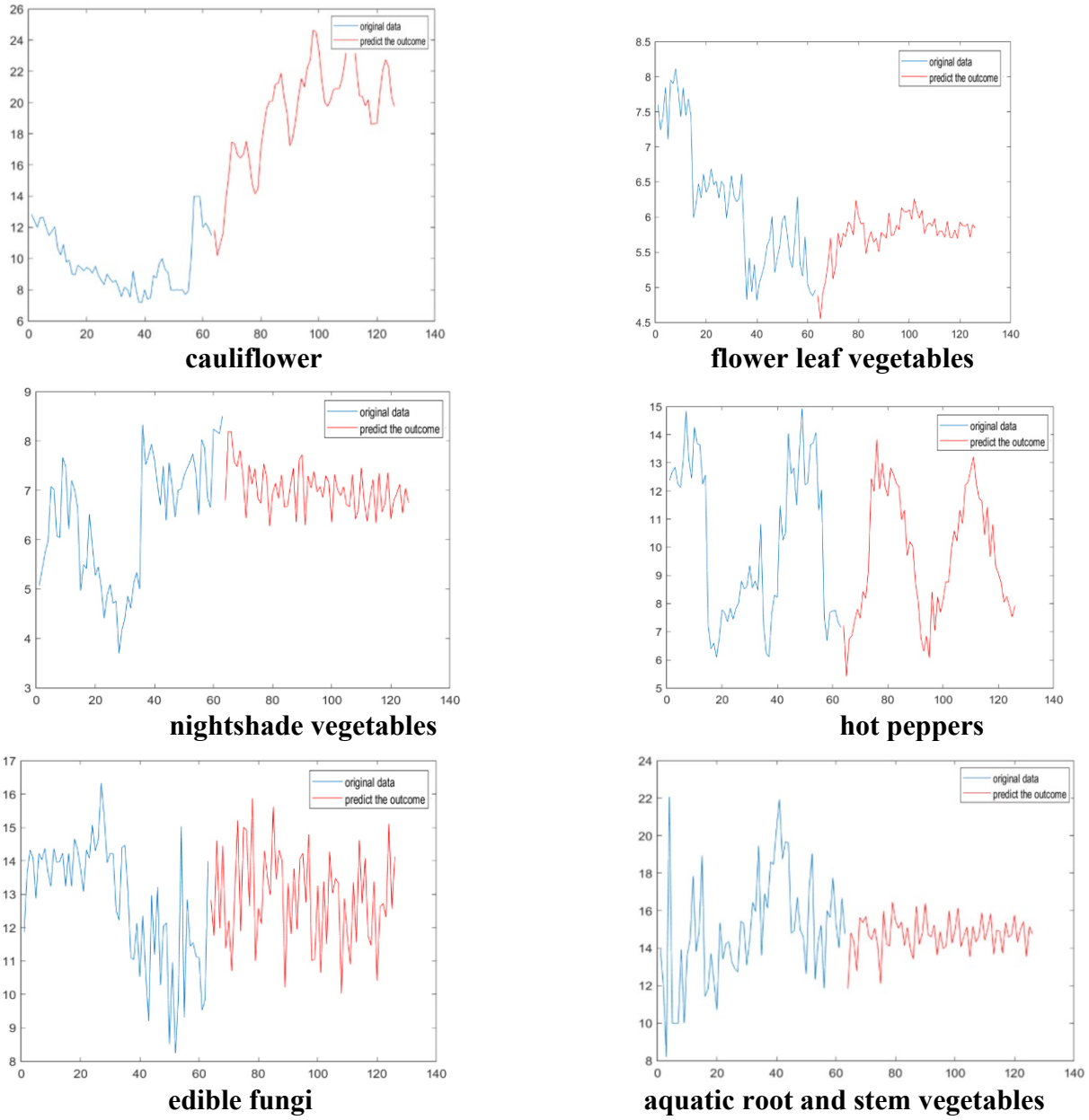


Figure 2 Prediction of pricing for six categories of vegetables.

Figure 2 represents the predicted pricing for vegetables over a certain period of time, with the blue line indicating the known pricing and the red line indicating the predicted pricing. The study took into account various factors that influence vegetable pricing, resulting in reasonable results. By analyzing the predicted curve, pricing strategies for vegetables from July 1st to July 7th, 2023 were determined. Table 2 shows the specific pricing for the six categories of vegetables.

Table 2 Prediction of Pricing for Six Categories of Vegetables on July 7th (unit: Yuan)

	flower leaf vegetables	cauliflower	aquatic root and stem vegetables	nightshade vegetables	hot peppers	edible fungi
July 1st	4.886	11.822	11.826	6.780	7,226	12.834
July 2st	4.549	10.192	14.808	8.190	5.416	11.761
July 3st	4.929	10.902	14.350	8.188	6.741	14.611
July 4st	5.073	11.633	12.790	7.578	6.866	11.982
July 5st	5.341	13.838	15.639	7.479	7.390	14.455
July 6st	5.703	15.395	15.348	7.813	7.800	11.387
July 7st	5.118	17.455	15.685	7.261	7.472	12.206

4. Conclusions

A linear regression model was applied to explore the relationship between sales volume and cost markup pricing for six categories of vegetables offered for sale in a supermarket. It was found out that an insignificant correlation was present between cost markup pricing and sales volume. Then, a time series ARIMA forecasting model was adopted to make pricing decisions on various categories in the first week of July 2023 in the supermarket. The ARIMA model was used to perform an Augmented Dickey-Fuller (ADF) test for the stability of the time series data to be assessed. As for the relatively unstable data, it is necessary to input a large amount of data for the maximum accuracy in predicting the future trend data development wherever possible.

In practice, time series data often exhibit clear periodic patterns. This study involves the seasonality of vegetables in the supermarket, and a large volume of data is available for the research object. Therefore, it is a reasonable decision to predict the pricing of vegetable products in the supermarket using a time series forecasting model. This model is applicable to predict the periodic changes and development trends of time series data accurately, thus providing effective data support. With clear statistical meanings, its parameters can be effectively analyzed.

References

[1] Peng J. Latest Advances in Preservation Technology for Fresh Fruit and Vegetables [J]. Foods,2022,11(20).

[2] Suci M,Noor V H. Quantitative analysis on the fluctuation of vegetable price in supermarket [J]. MATEC Web of Conferences,2018,154.

[3] A Qi. Research on Pricing of Science and Technology Information Service Based on Cost-Plus Pricing Method [J]. Library Research and Work, 2021(10): 25-31+24.

[4] Shen, G., Wang, L., & Jiang, Y. (2022). Application of Multivariate Statistical Analysis Model in Product Market Demand Analysis: A Case Study of Demand Analysis for Hand Cleaning Products of Residents under the COVID-19 Pandemic. Commercial Exhibition Economy, 20, 79-81.

[5] Ya-Fang S,Yue-Jun Z,Bin S.How does global transport sector improves the missions reduction performance? A demand-side analysis [J]. Applied Energy2022,311.

[6] Oke MD,Lawanson IO,Muhammed AI.SUPPLY-SIDE ANALYSIS OF THE MARKET FOR TOMATO IN BAUCHI STATE,NIGERIA [J].The Journal of Developing Areas,2020,54(3).

[7] Walter D,Irene M,Averil G, et al. Linear regression equations to predict β-lactam, macrolide, lincosamide and fluoroquinolone minimum inhibitory concentrations from molecular antimicrobial resistance determinants in Streptococcus pneumoniae.[J]. Antimicrobial agents and chemotherapy,2021.

[8] Haokun W,Delin H,Maomao Z, et al. Multiphase flowrate measurement with time series sensing data and sequential model [J]. International Journal of Multiphase Flow,2022,146.

[9] Dong, Y. (2022). Construction and Research of ARIMA-BiLSTM Model in Vegetable Price Forecasting under Different Combination Methods (Doctoral dissertation, Dongbei University of Finance and Economics).

[10] Tian, X. (2023). Design of Carbon Futures Products Based on ARIMA Model (master's thesis, Anhui University of Finance and Economics).