A Study on Replenishment Decision of Vegetable Goods Based on Polynomial Regression and Arima Time Series Forecasting

Zikai Zhao *

School of Public Administration, Hohai University, Nanjing, China, 210000

* Corresponding Author Email: z1781262906@163.com

Abstract. In a fresh food superstore, the superstore replenishes daily based on the historical sales and demand of vegetable commodities. To formulate the optimal replenishment decision, this paper analyses the sales flow, purchase cost, and loss of a supermarket in a period, and through the use of a variety of models, it can give a reference opinion on the formulation of automatic pricing and replenishment decisions for vegetable commodities. The first text describes the process of analyzing the relationship between total sales volume and cost-plus pricing, including data integration, preprocessing, and multiple regression analysis methods. Through these analyses, a general relationship between total sales volume and merchandise selling price is identified. Univariate linear regression, polynomial regression, random forest regression, and BP neural network regression models were all used to analyze this relationship. The focus is then on the total daily replenishment and pricing strategy for the coming week to maximize the superstore's revenue. The text details the use of ARIMA time series analysis methods, as well as steps for noise reduction and forecasting using Wavelet wavelet functions. Through these methods, the text provides sales forecast results for different vegetable categories for the coming week for the superstore management's reference.

Keywords: Regression model; ARIMA time series; Wavelet filter function; vegetable commodity replenishment decisions.

1. Introduction

In the daily sales of fresh food supermarkets, the shelf life of fresh food is exceptionally short, creating a constant need for replenishment to meet customer demands and maintain product freshness [1-2]. With a wide range of fresh food items being sold, and sourced from various origins, the bustling activity of vegetable trading begins during the early hours of the day, typically taking place between 3:00 and 4:00 a.m. It is fascinating to observe that during this crucial replenishment decision-making process, the businessmen involved are not privy to the specific assortment of fresh food items or their corresponding purchase prices [3]. As a solution, supermarkets often resort to implementing a "cost-plus pricing" strategy, enabling them to offer discounts on damaged or subpar products while ensuring profitability. Given the dynamic nature of the market, gaining accurate insights into consumer demand becomes paramount for supermarkets to make well-informed decisions concerning replenishment and pricing strategies. Building upon previous research, noteworthy correlations have been identified, linking the sales volume of vegetable items with the time of day, as well as their corresponding prices and overall sales quantities [4-5].

The primary objective of this paper is to develop comprehensive replenishment plans, incorporating a category-by-category approach specifically tailored for superstores. Through an in-depth analysis of the intrinsic relationship between the total sales volume of each vegetable category and the utilization of cost-plus pricing, we aim to provide invaluable guidance on determining the optimal daily replenishment volume and pricing strategy for each vegetable category. The proposed research focuses on studying the period spanning July 1-7, 2023, with the ultimate goal of maximizing the superstore's overall profit margin. As supermarkets continue their pursuit of operational excellence, this study holds the potential to yield valuable insights that can be applied in the development and implementation of automated pricing and replenishment decision-making systems within the realm of vegetable goods. It is worth noting that the extensive data utilized in this paper has been sourced from the distinguished 2023 National College Students Mathematical Modeling Competition Question C [6].

2. Modeling and solving the relationship between sales volume and pricing

2.1. Univariate linear regression

With the sales volume of each vegetable category as the independent variable and cost-plus pricing as the dependent variable, after univariate linear regression analyses of the pre-processed data, the results are obtained in Figure 1 below:

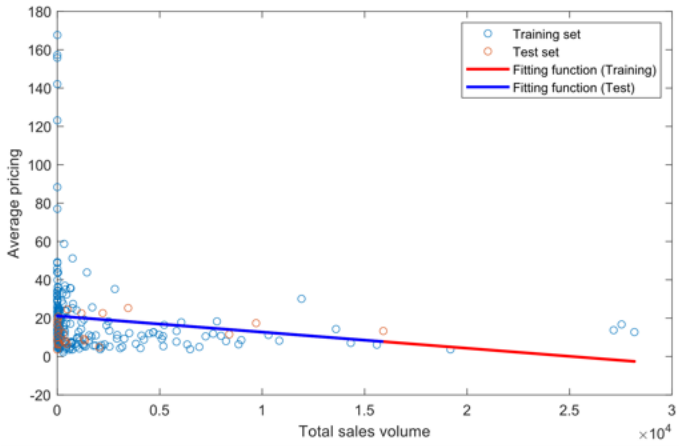

Figure 1. Linear regression analysis results.

The linear regression equation obtained for total sales and average pricing is:

$$y = -0.00084178222720775277x + 21.14324 \quad (1)$$

The above model, which was trained and tested using a training set to test set ratio of 9:1, has the following error metrics in Table 1.

Table 1. Error Metrics.

	MAP	RMSE	MAE	MAPE	R^2
Training Set	544.8225	23.3414	13.145	113.3617%	0.022825
Test Set	97.2206	9.86	8.4988	110.1606%	-1.3307

Evaluate the regression results of poor fit, but we can univariate through the linear regression results found that the slope of the regression equation is negative[7], the total sales volume and the selling price of goods between the negative correlation, which is in line with the general law of supply and demand in microeconomics;

2.2. Perform polynomial regression analyses

The binomial and trinomial regressions were implemented using Matlab and the fitted images and error metrics are shown in Figure 2 and Table 2 below:

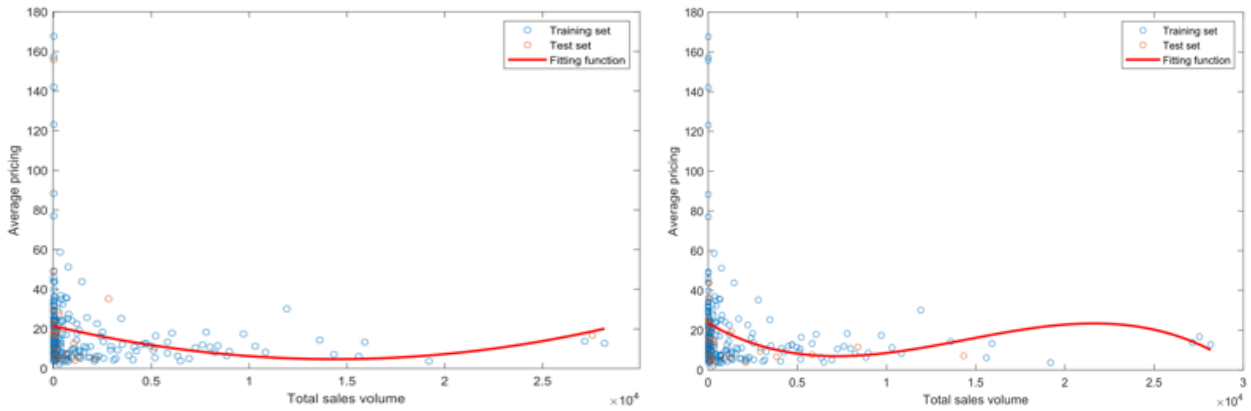

Figure 2. Results of binomial and cubic regression analyses.

Table 2. Error measures.

	MAP	RMSE	MAE	MAPE	R^2
Training Set (secondary)	448.2429	21.1717	11.8605	103.6805%	0.042505
Test Set (secondary)	843.9931	29.0516	13.8894	101.5998%	0.015626
Training Set (three times)	544.8225	23.3414	13.145	113.3617%	0.022825
Test Set (three times)	97.2206	9.86	8.4988	110.1606%	-1.3307

The regression analyses conducted on binomials and trinomials unveiled an intriguing trend: the R-squared values steadily increased, indicating an improved fit for the models. This observation hinted at the potential for even better fits with polynomial regression analyses featuring four or more terms. However, to our dismay, there was an unexpected complication when attempting to compile the Matlab code for such analyses. An error emerged, causing a loss in the matrix rank. This error hindered the successful execution of the regression and prevented us from delving deeper into the intricacies of polynomial regressions. Regrettably, due to the software limitations encountered, we were unable to pursue further studies employing more elaborate polynomial regression models. Despite this setback, the insights gained from our analyses of binomials and trinomials lay a foundation for future investigations and continue to contribute to the ever-evolving field of regression analysis.

2.3. Random Forest Regression Analysis

The steps of random forest regression analysis are:

Step1 collecting relevant data and pre-processing the data including handling missing values, standardization or normalization, and other manipulations [8];

Step2Constructing a random forest regression model and training each decision tree using a randomly selected sample of independent variables and a randomly selected sample of observations;

Step3 obtain the final regression prediction results by averaging or voting the prediction results of multiple decision trees;

Step4 uses the dataset for training the random forest regression model, adjusting the number of decision trees and other parameters;

Step5 Assessing the performance of the random forest model by error metrics using methods such as cross-validation;

Step 6 use the trained random forest regression model for prediction and inference of the dependent variable.

Random forest regression analysis was carried out through SPSSPRO and the fitting results and error assessment values are shown in Figure 3 and Table 3 below:

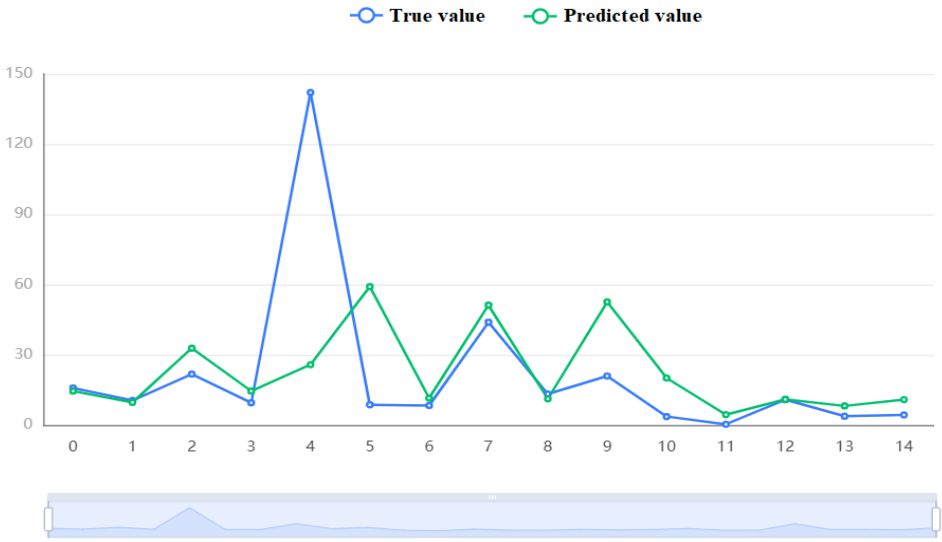

Figure 3. Fitting results.

Table 3. Error assessment values.

	MAP	RMSE	MAE	MAPE	R^2
Training Set	140.202	11.841	6.703	34.201	0.706
Test Set	883.455	29.723	16.518	72.648	-0.203

Random forest regression showed a significant improvement in fit compared to monomial and polynomial regression;

2.4. BP-neural network regression analysis

BP-neural network regression was performed using the Neural Network Toolbox library in Matlab, using the feedforward net function to create a feedforward neural network model; the trainFcn parameter was used to set the training function; the divider and function were used to perform random division; the train function was used to train the neural network model Use sim function to predict the test set and get the output of the model [9].

The optimal solution of the 11th round is obtained when the number of cycles of training = 17, and the fitting results of the model are shown in Fig. 4:

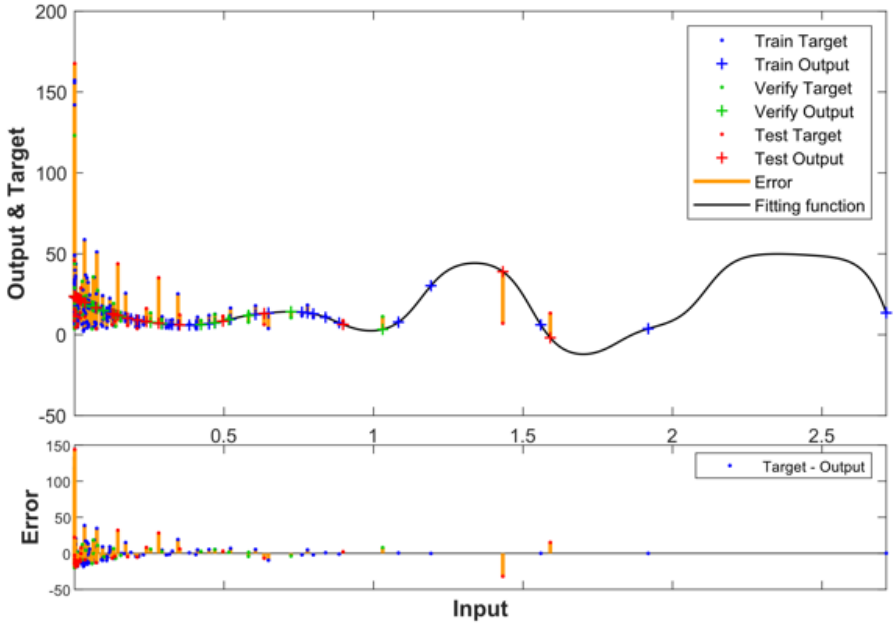

Figure 4. Fitting results of the BP neural network model.

After using the BP-neural network deep learning algorithm, the fit of the curves has been significantly increased [10].

Different regression models possess different strengths. Using the linear regression model and polynomial fitting, although the fit is not high, we can find the general relationship between sales volume and the selling price of goods through the fitted curve, which can provide a reference basis for the superstore manager for the pricing of dishes and the number of goods. The machine learning model has a higher degree of fit, and for a given pricing of a new dish, it can predict its future sales.

3. Establishment and solution of replenishment and pricing model

3.1. Forecasting using ARIMA time series

ARIMA (Autoregressive Integrated Moving Average) is a commonly used time series analysis method for building forecasting models for time series data. The ARIMA model can handle time series data with autoregressive (AR) and moving average (MA) characteristics.

The autocorrelation coefficient (ACF) measures the degree of correlation between the same events over two different periods and is calculated as follows:

$$ACF(k) = \sum_{t=k+1}^{n} \frac{(Z_t - \bar{Z})(Z_{t-k} - \bar{Z})}{\sum_{t=1}^{n}(Z_t - \bar{Z})^2} \tag{2}$$

The partial autocorrelation coefficient (PACF) represents the degree to which the influence or correlation of one element is calculated by treating the influence of the other elements as constant, i.e., by disregarding the influence of the other elements and examining the degree of closeness of the interrelationship between the two elements alone. The PACF is calculated as follows:

$$\begin{aligned} PACF(k) &= \frac{E(Z_t - EZ_t)(Z_{t-k} - EZ_{t-k})}{\sqrt{(Z_t - EZ_t)^2}\sqrt{(Z_{t-k} - EZ_{t-k})^2}} \\ &= \frac{cov[(Z_t - \bar{Z}), (Z_{t-k} - \bar{Z})]}{\sqrt{var((Z_t - \bar{Z}_t))}\sqrt{var((Z_{t-k} - \bar{Z}_{t-k}))}} \end{aligned} \tag{3}$$

The AIC criterion is the minimization of the information criterion and is calculated as follows:

$$AIC = -2\ln(L) + 2K \tag{4}$$

When the sample size is large, the BIC Bayesian Information Criterion is used:

$$BIC = -2\ln(L) + Kln(n) \tag{5}$$

Where L denotes the great likelihood function of the model, K denotes the number of model parameters, and n denotes the sample capacity, by comparing the values of AIC and BIC with different difference orders and taking the smallest value of the two, p and q.

In this paper, we take a daily sales volume of large data - amaranth as a model cited example, and finally, we automatically find the optimal parameters based on the AIC information criterion, the model results for the ARIMA model (1, 0, 3) test table, based on the variables: commodity 102900005115762_min-max standardization, from the Q The analysis of the results of the statistics can be obtained: Q6 does not show significance in the level, cannot reject the hypothesis that the residuals of the model are white noise series, at the same time, the model's goodness of fit R^2 is 0.679, the model performs better, the model meets the requirements.

The model formula is as follows:

$$y(t) = 0.138 + 0.987y(t-1) - 0.557\varepsilon(t-1) - 0.186\varepsilon(t-2) - 0.025\varepsilon(t-3) \tag{6}$$

After the inverse normalization process, the fitted image is derived with the prediction as shown in Figure 5 below:

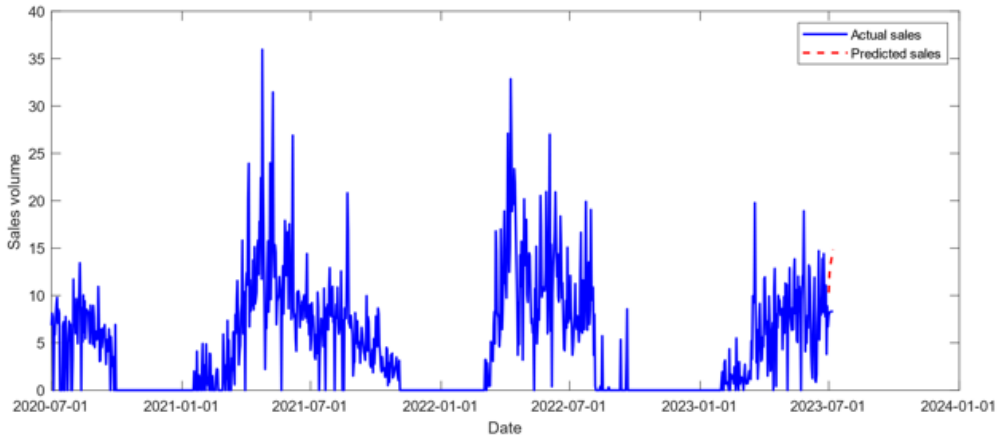

Figure. 5 Fitted image and prediction

3.2. Model Optimisation using Wavelet Function

In the ARIMA time series model of "Amaranthus", we found from the analysis of the Q statistic results that Q6 does not show significance at the level, and we cannot reject the hypothesis that the residuals of the model are white noise series. Therefore it is necessary to introduce a noise reduction

model. In this paper, the Wavelet Transform function is introduced for noise reduction of the initial data.

Wavelet Transform has two styles: continuous Wavelet Transform (Continuous Wavelet Transform) and discrete Wavelet Transform (Discrete Wavelet Transform). Mathematically, the Continuous Wavelet Transform can be expressed by the following equation:

$$X_w(a,b) = \frac{1}{\sqrt{|a|}} \int x(t)\overline{\varphi}\left(\frac{t-b}{a}\right) dt \tag{7}$$

Where φ (t) is the continuous mother wave, a is the scale factor and b is the translation factor. The values of the scale factor and translation factor are continuous. In Matlab, you can use the Wavelet Toolbox library to perform wavelet analysis on the Amaranth data: use the wavedec function to perform wavelet decomposition, and decompose the signal into wavelet coefficients of different scales and frequencies. You can specify the number of decomposition layers and wavelet functions; use the splotches function to visualize the wavelet coefficients and observe the spectral characteristics of the signal at different scales; use the wavered function to reconstruct the wavelet coefficients and get the reconstructed signal; you can use compress function for wavelet compression.

The noise points are plotted using the plot function, and the horizontal axis indicates the number of days from 2020-7-1, see Figure 6:

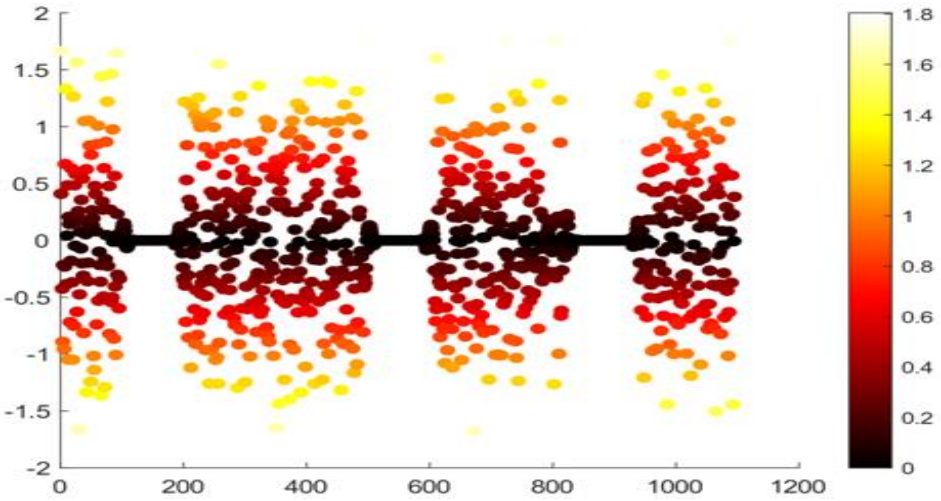

Figure 6. Noise Points.

After wavelet noise reduction, export the comparison between before and after noise reduction Figure 7:

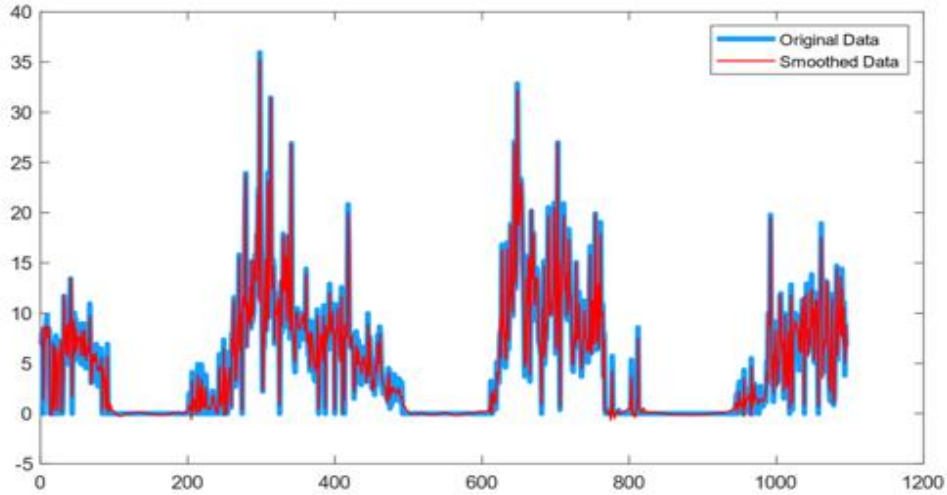

Figure 7. Comparison graph before and after noise reduction.

Predictions are made on the noise-reduced data and the results are shown in Figure 8:

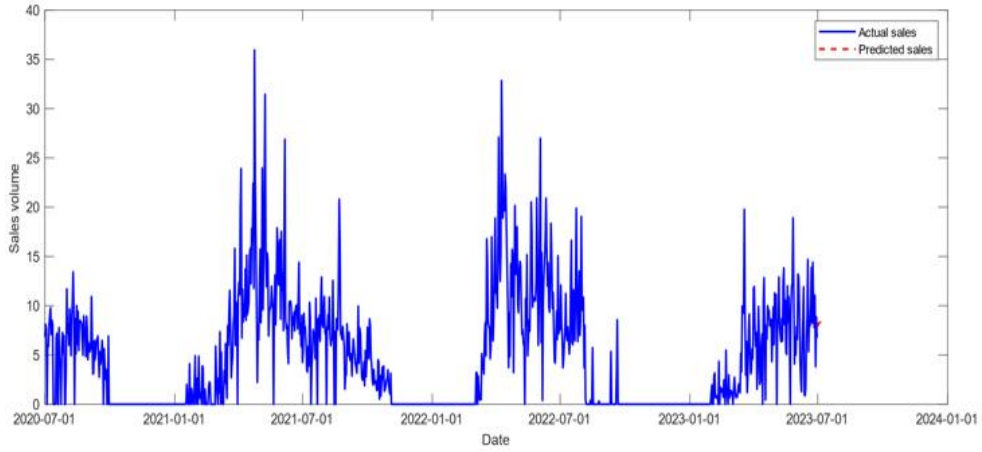

Figure 8. Prediction of data after noise reduction.

The sales volume of Amaranth for the next seven days is shown in Table 4 below:

Table 4. Sales of Amaranth for the next seven days.

Time	2023-7-1	2023-7-2	2023-7-3	2023-7-4	2023-7-5	2023-7-6	2023-7-7
Volume (kg)	7.8684	8.1851	8.2738	8.2996	8.3081	8.3117	8.3140

The for loop in Matlab is used to reduce the noise and predict the data for each dish. To ensure the authenticity of the results, the exported results treat all data with predicted sales less than 0.001 or negative values as zero. A more reliable daily replenishment total for each vegetable category for the coming week (1-7 July 2023) was derived.

4. Conclusions

This research paper aims to address the complex challenge of making replenishment decisions and pricing strategies for vegetable commodities in fresh food supermarkets. The nature of vegetable commodities, with their short shelf life and need for prompt replenishment based on historical sales and demand, adds to the intricacy of this task. The study, conducted by researchers from two universities in Beijing and Changsha, endeavors to offer informative recommendations for automated pricing and replenishment decisions specifically tailored to vegetable commodities. The investigation begins by establishing the relationship between total sales volume and the utilization of cost-plus pricing. To achieve this, the researchers undertake various data integration, preprocessing, and regression analysis methods. Notably, univariate linear regression, polynomial regression, random forest regression, and BP neural network regression are employed. These diverse regression models allow for a comprehensive examination of the relationship between total sales volume and the selling price of goods. Interestingly, the analyses reveal a negative correlation between these variables, aligning with the fundamental principles of supply and demand in microeconomics. Moreover, the different regression models employed in the study offer valuable insights, enabling the superstores to make informed decisions regarding dish pricing and stocking quantities.

Moving forward, the research extends its focus to encompass the total daily replenishment and pricing strategy for the upcoming week, with the overarching goal of maximizing the superstore's revenue. In this realm, ARIMA time series analysis and Wavelet Function Noise Reduction methodologies take center stage. The ARIMA model is leveraged to forecast sales data, while the Wavelet Function Noise Reduction technique is utilized to enhance the accuracy of these predictions. By incorporating these sophisticated methods, the study furnishes sales forecasting outcomes for various vegetable categories for the upcoming week. These results prove instrumental in providing robust decision support to superstore management, enabling them to optimize their operations and strategy. Through an integrated approach that combines regression analysis, time series forecasting, and noise reduction techniques, this research paper illuminates key insights into the replenishment

decisions and pricing strategies for vegetable commodities in fresh food supermarkets. By bridging the gap between theory and practical applications, this study contributes to the existing knowledge base and empowers superstores with valuable guidance in this complex and ever-evolving domain.

References

[1] PAN Xiaofei, XIE Zhiheng, WANG Shuyun. Optimal decision-making of fresh food superstore preservation efforts and pricing considering loss aversion [J]. Highway Traffic Science and Technology, 2022, 39(06):177-185+190.

[2] WANG Yongsheng. Research on enterprise value creation path based on core competitiveness evolution [D]. Xi'an University of Technology, 2023.

[3] Du Ji-liang, Hou J, Wu Shangyu et al. Evaluation model of discount strength in shopping malls [J]. Science and Technology Wind, 2019(32):235.

[4] Li Xiaolu, Zhou Shuguang. Research on the Development Problems of China's Fresh Food Supermarket Retail Industry [J]. Commercial Economy Research, 2021(23):35-37.

[5] Ren H.R.,Zhu Y.Y.. Research on data pricing model based on data market type [J]. Big Data, 2023, 9(4):116-138. DOI:10.11959/j.issn.2096-0271.2023052.

[6] Deepa R, Pradeep M, Soumik R, et al. Modelling of rainfall time series using NAR and ARIMA model over western Himalaya, India[J]. Arabian Journal of Geosciences, 2022, 15(23).

[7] Arfaoui S, Mabrouk B A, Cattani C. Wavelet Analysis: Basic Concepts and Applications [M].CRC Press: 2021-01-12.

[8] Zeng St. Exploration on the operation of township superstores under the background of new urbanization [J]. Chinese and foreign entrepreneurs, 2015(08):44-45.

[9] WANG Mengwei, ZHOU Yue, BAI Li et al. Early warning analysis of price fluctuation of spicy vegetables in China [J/OL]. China Melon and Vegetable: 1-16[2023-10-25].

[10] Yanjun Hu, Pingchuan Zhang, Zheng Shang et al. Research on garlic price prediction based on deep learning [J]. Journal of Henan Institute of Science and Technology (Natural Science Edition), 2023, 51(03):35-42.