

Vegetable Pricing and Replenishment Strategy Based on Multivariate Polynomial Regression and SARIMA Model

Tianrun Yu[#], Junhao Pu[#], Jiaming Guo[#], Qidan Guo, Qian Zhang^{*}

The College of Information, Mechanical and Electrical Engineering, Shanghai Normal University
Shanghai, China, 201400

^{*} Corresponding Author Email: qianzhang@shnu.edu.cn

[#]These authors contributed equally.

Abstract. Fresh products are one of the important parts of the profitability of superstores, and it is very important for merchants to make effective replenishment and pricing suggestions based on historical data. In order to give a reasonable vegetable pricing and replenishment plan, this paper processes the data and then uses multivariate polynomial regression to regress the sales volume and cost-plus pricing of the category, to obtain the relationship between the sales volume of the category and the cost-plus pricing; and combines with the SARIMA model to further give the recommended replenishment quantity; considering the impact of vegetable freshness on pricing, a freshness-based pricing model is proposed, and the optimal pricing and optimal restocking quantity formula, and proved its correctness in principle. In this study, the pricing and replenishment strategies of reference vegetables are given, and future research can further explore the market forecast of other agricultural products on this basis to broaden the application scope of this model.

Keywords: SARIMA, Multivariate Polynomial Regression, Vegetable Pricing.

1. Introduction

In the current global agricultural market, accurate forecasting of vegetable sales volumes is crucial for developing effective supply chain management strategies and pricing decisions. In particular, seasonal variations have a significant impact on vegetable sales volumes, which need to be scientifically managed and optimised by efficient forecasting models. In recent years, a variety of time series forecasting models, such as ARIMA, SARIMA, and composite models combined with modern machine learning techniques, have been widely studied and applied in this field.

From the use of SARIMA model in predicting the monthly rainfall frequency in Umuahia [1] to its application to the number of boat passengers in Batam City [2], SARIMA has been shown to be highly efficient in dealing with seasonal and cyclical data. In addition, the application of the SARIMA model augmented by wavelet transform in agricultural price forecasting [3] further demonstrates the model's ability to handle complex time series data. In an application example in Nairobi, Kenya [4], the SARIMA model effectively supported market pricing decisions with its accurate forecasting performance.

The study of cauliflower price forecasting in Nepal [5] and the forecasting of agricultural commodity prices in Mexico [6] further validate the wide applicability of time series forecasting models in price forecasting of agricultural products in different regions of the world. KY Choong also emphasised the importance of the SARIMA model in the decision making process [7]. By analysing and predicting the price fluctuations of these key crops, it can provide local farmers with important market entry information to help them optimise their planting plans and marketing strategies.

In addition, some studies have continuously improved the accuracy and usefulness of prediction models by introducing genetic algorithms [8] and random forest models combined with reinforcement learning techniques [9]. These innovative approaches not only enhance the predictive efficiency of the models, but also make the models more adaptable to the changing market conditions.

Through the combined application of multivariate polynomial regression, SARIMA and its related extended models, we can not only predict the seasonal sales changes of vegetables and other

agricultural products more accurately, but also provide merchants with replenishment and pricing strategies with reference value.

Multivariate polynomial regression is a crucial technique used in this study to model the relationship between multiple independent variables and the dependent variable, specifically the sales volume and cost-plus pricing of various vegetable categories. This method allows for capturing non-linear relationships that are often present in market dynamics. By using multivariate polynomial regression, we can establish a comprehensive model that integrates various factors influencing sales and pricing, providing a more detailed and accurate analysis. This technique is particularly valuable as it complements the SARIMA model, which primarily handles time series data, by adding the ability to account for multiple interacting factors in the pricing and sales volume of vegetables.

2. Principles of replenishment strategies and pricing model

2.1. Replenishment strategy based on multivariate polynomial regression with SARIMA

(1) Category Pricing Model Based on Multivariate Polynomial Regression

Let the random variable y vary with m independent variables $x_1, x_2, \dots, x_k$, and assume the relationship is given by:

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \varepsilon \quad (1)$$

Where $\beta_0, \beta_1, \dots, \beta_k$ are called regression coefficients, which are $k+1$ parameters to be estimated, and ε is a random variable.

The primary problem in regression analysis is to estimate the regression coefficients β_i using n sets of observed data for $x_1, x_2, \dots, x_k, y$, denoted as $(x_{k1}, x_{k2}, \dots, x_{km}, y_k)$ for $k=1, 2, \dots, n$. Each β_i needs to be estimated as $\hat{\beta}_i$ and tested statistically. Substituting these estimates into the regression equation, we obtain:

$$\begin{cases} y = \beta_0 + \beta_1 x_{11} + \dots + \beta_m x_{1m} + \varepsilon_1 \\ y_n = \beta_0 + \beta_1 x_{n1} + \dots + \beta_m x_{nm} + \varepsilon_n \end{cases} \quad (2)$$

In this equation $\varepsilon_1, \dots, \varepsilon_n$ are n independently and identically distributed random variables, each following a normal distribution $N(0, \sigma^2)$ let?

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1x_{11} & \dots & x_{1m} \\ \vdots & \ddots & \vdots \\ 1x_{n1} & \dots & x_{nm} \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \quad (3)$$

Thus, the corresponding matrix equation can be expressed as:

$$Y = X\beta + \varepsilon \quad (4)$$

For the data in this study, multivariate polynomial regression was employed to determine the relationship between category pricing and sales volume, as well as cost-plus pricing, for cauliflower, leafy vegetables, mushrooms, peppers, eggplants, and aquatic root vegetables. The relationships are as follows. Since the coefficients of S_i^2 are less than 0.0005, they have been omitted for computational convenience:

$$y = 3.426 + 0.637 \cdot P_i + 0.084 \cdot S_i + 0.060 \cdot B_i^2 - 0.017 \cdot B_i \cdot S_i \quad (5)$$

$$y = 7.749 - 1.543 \cdot P_i - 0.020 \cdot S_i + 0.234 \cdot B_i^2 - 0.00519 \cdot B_i \cdot S_i \quad (6)$$

$$y = 2.162 + 0.411 \cdot P_i + 0.00692 \cdot S_i + 0.140 \cdot B_i^2 - 0.00201 \cdot B_i \cdot S_i \quad (7)$$

$$y = 5.031 + 0.586 \cdot P_i - 0.0300 \cdot S_i + 0.00835 \cdot B_i^2 + 0.00352 \cdot B_i \cdot S_i \quad (8)$$

$$y = 3.198 + 0.977 - 0.0642 \cdot S_i + 0.0151 \cdot B_i^2 + 0.00838 \cdot B_i \cdot S_i \quad (9)$$

$$y = -2.513 + 2.392 \cdot P_i + 0.00551 \cdot S_i - 0.0674 \cdot B_i^2 - 0.00191 \cdot B_i \cdot S_i \quad (10)$$

Where B_i represents the wholesale price, P_i represents the cost-plus price, and S_i represents the daily sales volume.

(2) Inventory Replenishment Model Based on SARIMA

In the SARIMA $(p, d, q)(P, D, Q)$ model structure, p represents the number of autoregressive terms, q represents the number of moving average terms, d represents the number of differences needed to make the time series stationary, P represents the seasonal autoregressive order, Q represents the seasonal moving average order, and D represents the number of seasonal differences. Here, s denotes the length of the seasonal cycle.

The specific steps are as follows:

Decompose the original time series into trend data, seasonal data, etc., to intuitively assess the seasonal effects in the data.

Perform a stationarity test on the time series. If $P < 0.05$, the series is stationary. If the original time series does not meet the stationarity requirement, apply differencing and seasonal differencing until the series becomes stationary.

Conduct partial autocorrelation (PACF) analysis on the processed time series to estimate the values of p and q .

To ensure the ARIMA model's residuals are white noise (i.e., the model is purely random), check the Q -statistic P -value ($P > 0.05$ indicates white noise). Additionally, analyze the information criteria values AIC and BIC.

The profit function is as follows:

$$f_i = s_i \times (p_i - c_i) \quad (11)$$

Where c_i the weighted purchase is cost of cauliflower on day i ; p_i is the selling price on day i ; S_i is the sales volume on day i ; and f_i is the profit on day i .

By substituting the pricing formulas for each vegetable category into the category pricing model and combining them with equation (10), the final replenishment quantity can be determined.

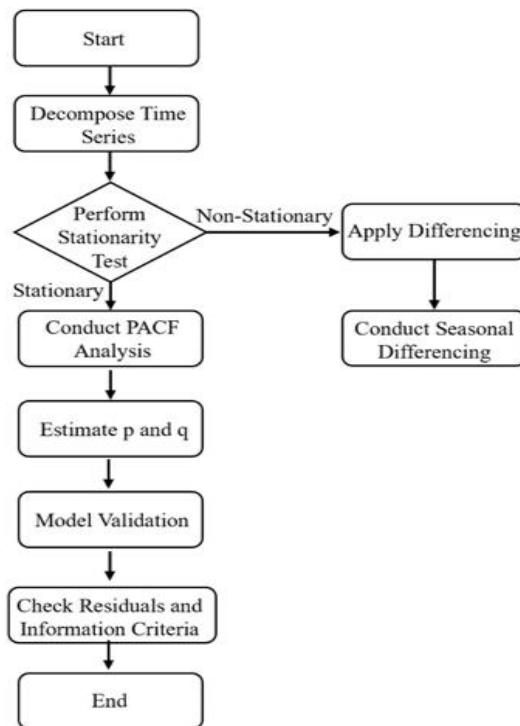


Figure 1. Flow Chart of Inventory Replenishment Model Based on SARIMA.

2.2. The Structure of the Freshness-Based Pricing Model

Most vegetables have a short shelf life, with almost all vegetable sales cycles lasting only one day. Prices typically change only twice within a day, leading to the assumption that one day represents the

entire sales cycle for all vegetables. Given the difficulty of quantifying freshness in the context of this study, this paper adopts the definition of the freshness function as provided in reference [10]:

$$\theta(t) = \frac{1}{1+\alpha t^2} \quad (12)$$

Where α represents the rate at which freshness declines over time.

Based on the basic form of the demand function, it can be assumed that the demand function for customers concerning vegetable products, as the freshness of vegetables continuously decreases, is given by:

$$D(p, t) = a\theta(t) - bp \quad (a > 0, b > 0) \quad (13)$$

Where a represents the basic market demand and b is the price elasticity factor. At this point, it can be considered that the demand function is not only a decreasing function with respect to price but also an increasing function with respect to freshness.

The order quantity can be regarded as the accumulation of sales volume at each micro-moment, that is:

$$Q = \int_0^T D(p, t) dt \quad (14)$$

In the absence of other profits, the profit is the difference between the total sales revenue and the total purchasing cost.

$$L(p, T) = \int_0^T pD(p, t) dt - Qw \quad (15)$$

Where Q is the total purchase quantity, w is the purchase price, T is the sales cycle (i.e., one day), and t is the time already elapsed in the sales cycle. From the analysis, it can be concluded that, to achieve maximum profit, the optimal price p_a must satisfy:

$$\frac{\partial L(p_a, T)}{\partial p_a} = 0 \quad (16)$$

The optimal decision problem can then be transformed into finding the optimal price $p^*(t)$ and the optimal order quantity Q^* that maximize the profit, which requires solving:

$$L(p^*, T) = \max_{p(t)} \int_0^T pD(p, t) dt - Qw \quad (17)$$

The optimal price can be determined as:

$$P_a^* = \frac{\int_0^T \alpha \theta(t) dt}{2bT} \quad (18)$$

Thus, the optimal order quantity is:

$$Q^* = \frac{\int_0^T \alpha \theta(t) dt}{2} \quad (19)$$

2.3. Verification of Maximum Profit

For the above conclusion, it can be demonstrated that dynamic pricing yields higher profits compared to a single fixed price. The proof process is as follows:

$$\Delta L = L(p^*(t), T) - L(p_a^*, T) = \frac{1}{4b} \int_0^T [a\theta(t)]^2 dt - [\int_0^T a\theta(t)]^2 \quad (20)$$

Let $h(t) = a\theta(t)$, then let:

$$F(T) = T[\int_0^T f^2(t) dt] - [\int_0^T f(t) dt]^2 \quad (21)$$

When $F(0) = 0$, the following relationship holds:

$$\frac{dF(T)}{dT} = \frac{1}{4bT} \int_0^T [h(t) - h(T)]^2 dt \geq 0 \quad (22)$$

Therefore, it is proved that for different vegetable items, the reference values of optimal pricing and optimal purchase quantity can be obtained by bringing the corresponding data into the above model. The determination of the number of nodes in the hidden layer is a very important and complex problem.

3. Results

3.1. The establishment of simulation model

This study collected vegetable sales information from July 1, 2020, to June 30, 2023. The dataset comprises 878,504 vegetable sales records, with 600,000 records used for training and 278,504 records used for testing. The data categories include sales date, vegetable name, sales volume, and sales unit price. All text data underwent preprocessing, which included removing outliers, converting date formats, and eliminating illegal characters. The dataset was evenly divided into training and testing sets, each accounting for 50% of the total data. The distribution of data in the training and testing sets was kept consistent to ensure the fairness of the model evaluation. This dataset contains only vegetable sales information and does not include any personal identification information, thus avoiding privacy concerns. All data comply with ethical standards for the use of public data.

3.2. Analysis of experimental results

3.2.1. Replenishment Quantity Prediction Results and Analysis.

Given the following constraints:

(1) $0 < p_i < 2483.88$, where 2483.88 is the upper limit of the total sales volume across all categories based on historical data.

(2) The sales volume of each vegetable must be at least greater than the historical minimum sales volume, $s_i > s_{i,min}$, and the daily weighted purchase cost of cauliflower is known from previous SARIMA forecasts.

Determine the optimal selling price and sales volume that maximize profit for each vegetable, as listed in Table 2. Similarly, calculate the maximum replenishment quantity for the remaining five categories, as shown in Table 1.

Table 1. Prediction of replenishment quantities statistics.

Category	Maximum Replenishment Quantity (kg)
Leafy Vegetables	1265.473
Solanaceous Vegetables	118.931
Pepper Varieties	604.231
Edible Fungi	511.136
Aquatic Rhizomes	296.792
Cauliflower	186.155

Table 2. Prediction of Daily Sales Volumes Statistics.

Category Date	Cauliflower	Aquatic Rhizomes	Pepper	Fungi	Solanaceous Vegetables	Leafy Vegetables
7.1	48.23	5.81	12.01	7.83	2.52	4.77
7.2	46.91	5.83	12.02	16.85	2.52	3.08
7.3	44.75	5.99	10.71	8.29	2.52	39.02
7.4	47.60	5.93	11.82	7.85	2.52	9.62
7.5	47.84	5.82	11.45	9.72	2.52	41.28
7.6	46.30	5.90	11.96	8.30	2.52	43.04
7.7	50.01	5.84	12.01	10.77	2.52	28.97

The spoilage value can be derived from the SARIMA forecast, as shown in Table 3:

Table 3. Prediction of Spoilage Statistics for Each Category.

Category	Spoilage Statistics
Cauliflower	0.094
Leafy Vegetables	0.099
Solanaceous Vegetables	0.0602
Fungi	0.027
Pepper Varieties	0.0868
Aquatic Rhizomes	0.1127

Based on the formula Replenishment Quantity = Sales Volume / (1 - Spoilage Rate), the replenishment quantity can be determined as shown in Table 4:

Table 4. Prediction of Replenishment Quantities Statistics.

Category Date	Cauliflower	Aquatic Rhizomes	Pepper	Fungi	Solanaceous Vegetables	Leafy Vegetables
7.1	52.23	6.45	12.78	8.05	2.76	5.38
7.2	51.78	6.47	12.79	17.32	2.76	3.47
7.3	49.39	6.65	11.40	8.52	2.76	43.98
7.4	52.54	6.58	12.58	8.07	2.76	10.84
7.5	52.90	6.46	12.18	9.99	2.76	46.52
7.6	51.10	6.55	12.73	8.53	2.76	48.51
7.7	55.20	6.48	12.78	11.07	2.76	32.65

Table 5. MSE and R^2 for Each Vegetable Category.

	Cauliflower	Leafy Vegetables	Fungi	Pepper	Solanaceous Vegetables	Aquatic Rhizomes
MSE	0.16	0.644	0.131	0.35	0.585	0.495
R^2	0.689	0.815	0.925	0.749	0.896	0.958

As shown in the table 5. The goodness of fit (R^2) of the multiple polynomial regression model for the six vegetable categories is all greater than 0.689, with the highest reaching 0.958. This indicates that our model has excellent fitting capability for the features involved.

3.2.2. Pricing Results and Analysis.

The prices of leafy vegetables exhibit noticeable seasonal and cyclical characteristics, and their freshness declines rapidly during the sales process, making them suitable for multiple price adjustments within a single day. Therefore, we have chosen leafy vegetables as the primary focus of this study, with α set to 0.03. According to the comparison between the model's predicted values and the actual values in Table 6, the results indicate that the model's maximum deviation rate is 11.20% and the minimum deviation rate is 0.15%. This high level of accuracy suggests that the model can serve as a reliable reference for pricing these vegetables.

Table 6. Model Prediction Accuracy for Leafy Vegetables.

Date	Actual Value	Fitted Value	Deviation Rate
2023-05-19	141.327	141.117	0.15%
2023-05-20	177.822	186.499	4.88%
2023-05-21	197.446	196.183	0.64%
2023-05-22	154.18	171.447	11.20%
2023-05-23	150.721	147.195	2.34%

To validate the model's sensitivity, K-Fold cross-validation was employed to estimate the performance of each model. Specifically, we increased the data in the model by 1% to test its sensitivity. As shown in the Table7, with a 1% adjustment, the average prediction deviation is very

small, and the standard deviation of the prediction deviation is also minimal. From the above data, it can be observed that the model exhibits strong robustness when the original data varies by 1%.

Table 7. Sensitivity Analysis of Multiple Polynomial Regression Model (1%).

	Cauliflowe r	Leafy Vegetables	Fungi	Peppe r	Solanaceous Vegetables	Aquatic Rhizomes
Average Prediction Deviation	0.071	0.033	0.06	0.047	0.081	0.086
Standard Deviation of Prediction Deviation	0.038	0.019	0.021	0.016	0.031	0.011

4. Conclusions

This study integrates multiple polynomial regression and SARIMA models to explore the issues of vegetable pricing and replenishment strategies, achieving in-depth analysis and effective prediction of vegetable market dynamics. The research demonstrates that combining these two models can effectively capture the seasonal variations in market prices and supply volumes, enhancing the accuracy and reliability of predictions.

In the empirical analysis, a large-scale dataset was utilized to verify the efficiency of the model in practical applications. The model not only accurately predicted market demand and price fluctuations for vegetables but also provided solid data support for formulating scientific replenishment strategies. Additionally, sensitivity analysis of the model further confirmed its stability and adaptability under different market conditions.

The model offers customized pricing and replenishment solutions for different types of vegetables, aiding businesses in making more scientific and reasonable decisions in a complex market environment. Future work could further improve the model's predictive performance and application scope by incorporating methods such as neural networks.

References

- [1] N. C. C., O. C. B., Maxwell O., et al. Comparing SANN and SARIMA for forecasting frequency of monthly rainfall in Umuahia[C]// Scientific African Conference, 2020: e0062.
- [2] Reza, W. Comparison of the accuracy of seasonal data prediction values using SARIMA and winter exponential smoothing on the number of ship passengers in Batam City[C]// 2023 International Conference on Statistical Analysis, 2023: 632-637.
- [3] Diop Mamadou Diéne, Sadefo Kamdem Jules. Multiscale Agricultural Commodities Forecasting Using Wavelet-SARIMA Process[C]// International Conference on Quantitative Economics, 2022: 1-40.
- [4] Robert Mathenge Mutwiri. Forecasting of Tomatoes Wholesale Prices of Nairobi in Kenya: Time Series Analysis Using Sarima Model[C]// International Conference on Statistical Distributions and Applications, 2019: 46-53.
- [5] Giri, A., Giri, V. R. Forecasting cauliflower prices in Nepal: a comparative analysis using seasonal time series and nonlinear models[C]// Cogent Food & Agriculture Symposium, 2024: 10(1).
- [6] Paredes-Garcia WJ, Ocampo-Velázquez RV, Torres-Pacheco I, Cedillo-Jiménez CA. Price Forecasting and Span Commercialization Opportunities for Mexican Agricultural Products[C]// International Agronomy Conference, 2019: 826.
- [7] Choong K Y, Sudin S, Raof R A A, et al. Application of SARIMA model in forecasting vegetable price in Malaysia: E-commerce platform domain[C]// AIP Conference Proceedings, AIP Publishing, 2024: 2898(1).
- [8] Guo, Y. Integrating genetic algorithm with ARIMA and reinforced random forest models to improve agriculture economy and yield forecasting[C]// Soft Computing Symposium, 2024: 1685-1706. <https://doi.org/10.1007/s00500-023-09516-8>

- [9] Mahmoud Sayed Agbo, H. Forecasting agricultural price volatility of some export crops in Egypt using ARIMA/GARCH model[C]// Review of Economics and Political Science Conference, 2023: 123-133.
<https://doi.org/10.1108/REPS-06-2022-0035>
- [10] Yi-Juan Chen. Multi-stage Dynamic Pricing Research on Fresh Food[C]// Hainan Normal University Academic Conference, 2015: MA thesis.