Research on Pricing and Replenishment Strategies for Supermarkets Based on Revenue Maximization

Zheng Lu [#, *], Yuanshuo Wang [#], Kefei Liu [#]

College of Information Science and Engineering, Hohai University, Changzhou, China, 213200

* Corresponding Author Email: 2162410309@hhu.edu.cn

[#]These authors contributed equally.

Abstract. Based on the perspectives of time series and the supply chain, this paper investigates pricing and replenishment strategies for fresh supermarkets with the aim of maximizing profits. Through visual analysis, the vegetable category is thoroughly examined, revealing seasonal sales patterns. Furthermore, the correlation between different categories is explored using ADF tests and the DPCCA model. To optimize the replenishment quantity for the upcoming week, sales data from the past month is employed for nonlinear fitting. Assuming a fixed purchase price, the ARIMA model is utilized to forecast pricing, and optimal pricing and daily replenishment strategies for the next seven days are formulated through nonlinear programming methods.

Keywords: Time Series, DPCCA, Pricing Strategy, ARIMA, Supply Chain.

1. Introduction

In the field of fresh supermarket operations, reasonable pricing and replenishment strategies are crucial for maximizing revenue. In recent years, numerous scholars have conducted in-depth research on this topic. For instance, Yan Zheng [1] (2019) and Zhonglin Li [2] (2023) verified the effectiveness of time series analysis in sales and demand forecasting, while Yunhao Cui[3](2023) improved forecast accuracy using the CNN-PSO-LSTM model. Meanwhile, Yixuan Chen [4] (2021) and Jiaojiao Li et al [5] (2023) explored the optimization of fresh agricultural products from a supply chain perspective. However, existing research tends to focus on either pricing or replenishment individually, lacking comprehensive studies that combine the two. This paper aims to delve into pricing and replenishment strategies for fresh supermarkets based on time series analysis and supply chain management. Unlike previous research, this paper not only employs the DPCCA model to more accurately analyze time series correlations but also considers external factors such as the impact of the pandemic on sales distribution. Additionally, it explores the relationship between pricing and sales volume and utilizes the ARIMA model for price forecasting, striving to provide more precise operational strategies. (Data source: http://www.mcm.edu.cn/html_cn/node/c74d72127066f510a5723a94b5323a26.html)

2. Research on the distribution pattern among various vegetable categories

2.1. Exploration of the overall distribution pattern

Based on the preprocessed sample data, the overall sales status of various vegetable categories is analyzed, as shown in Figure 1, which visually visualizes the sales distribution over time.

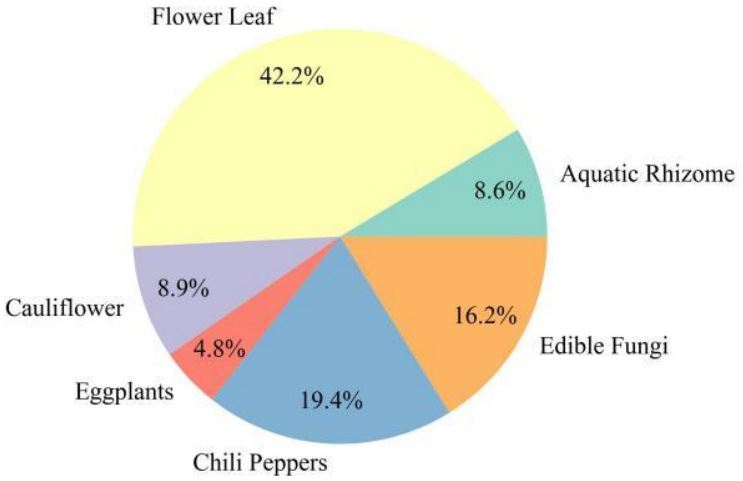

Fig 1. Overall sales distribution of various vegetable categories.

As shown in Figure 1, there is a significant difference in the sales volume of vegetable categories. The highest proportion is in the category of flowers and leaves, accounting for 42.2%, followed by chili peppers and edible fungi. The sales volume of aquatic roots and cauliflower is similar, while eggplants have the lowest sales volume, accounting for only 4.8%.

2.2. Trend analysis of seasonal distribution patterns

Then, considering the close relationship between the sales volume of various vegetable categories and the seasons [6], in order to further investigate the distribution of vegetable categories, this paper takes quarters as nodes, calculates the sales trends of different vegetable categories in each quarter, and draws a trend line chart, as shown in Figure 2.

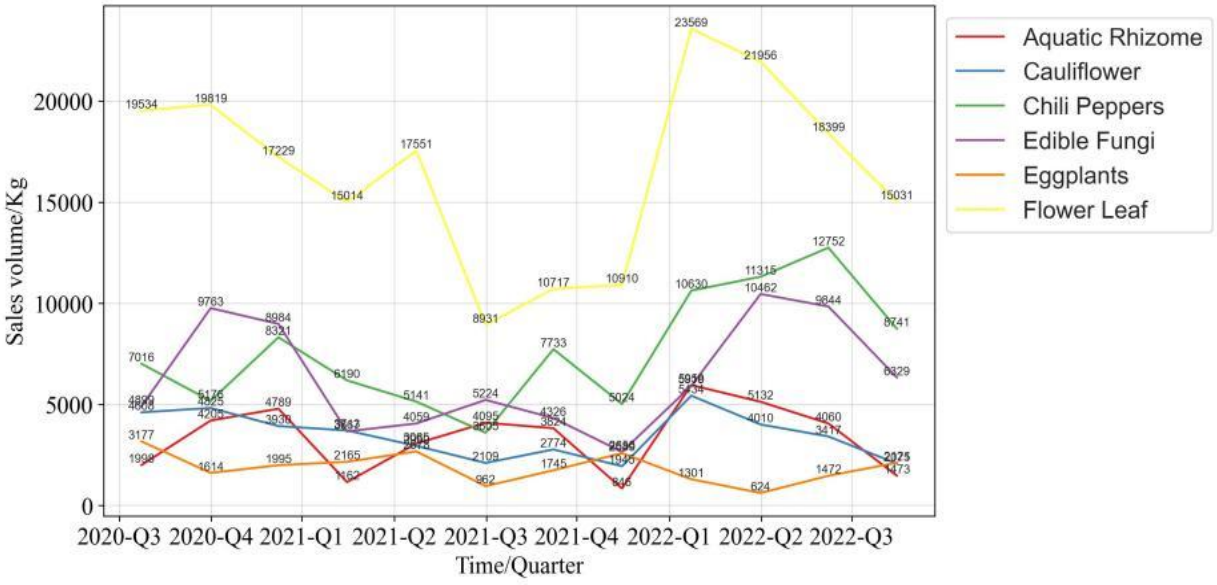

Fig 2. The trend of sales volume of various vegetable categories varying with seasons.

As shown in Figure 2, there are seasonal fluctuations in vegetable sales, with declines in the first and second quarters and an upward trend in the third and fourth quarters. This variation may be closely related to seasonal vegetable supply and demand, influenced by climatic conditions such as suitable winter growth and summer heat and rain. Additionally, year-end Spring Festival shopping and post-pandemic policy relaxation significantly boost sales. Prices, as a direct reflection of market supply and demand, are crucial to understanding sales distribution. As Figure 3 illustrates, a deeper exploration of vegetable prices across categories aids in further analyzing sales distribution patterns.

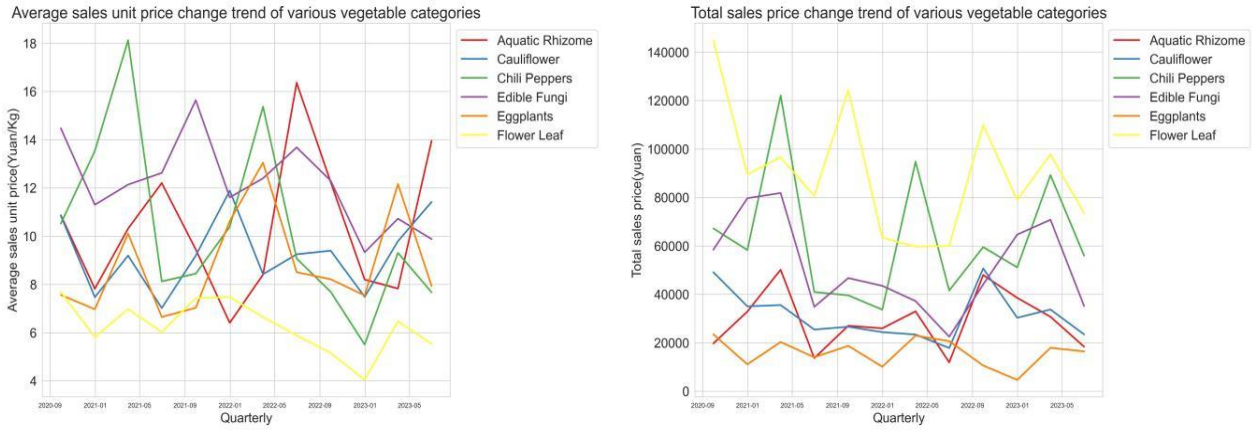

Fig 3. The trend of changes in unit sales price and total sales price for various vegetable categories.

According to Figure 3, despite having a relatively low unit price among various categories, flowering and leafy vegetables significantly lead in total sales price. After a thorough analysis of the sales distribution charts for each category, it was found that low unit pricing can significantly stimulate consumers' willingness to purchase, thereby driving an increase in sales volume. This discovery provides a new perspective for further understanding the distribution pattern of sales volume across categories.

2.3. Distribution of sales volume from the perspective of supply chain

2.3.1. Analysis of Supply Chain Solutions for Fresh Supermarkets

The supply chain is a functional network encompassing procurement of raw materials to consumer product sales. The vegetable supply chain in supermarkets often adopts the "farm-supermarket direct connection" model, seamlessly connecting farmers, businesses, and consumers for smooth vegetable production and sales. Large farmers play a pivotal role in this model, facilitating efficient information and fund circulation within the supply chain. When purchasing from producers, sales ends often use cost-plus pricing to ensure profits for all parties and protect consumer interests.

2.3.2. Analyzing and interpreting the distribution pattern of sales volume based on time series

Based on the aforementioned analysis, time series data often exhibit periodic variations. This paper aligns the supply chain sales phase with the epidemic spread stage to further elucidate the distribution pattern of sales volume and its underlying causes.

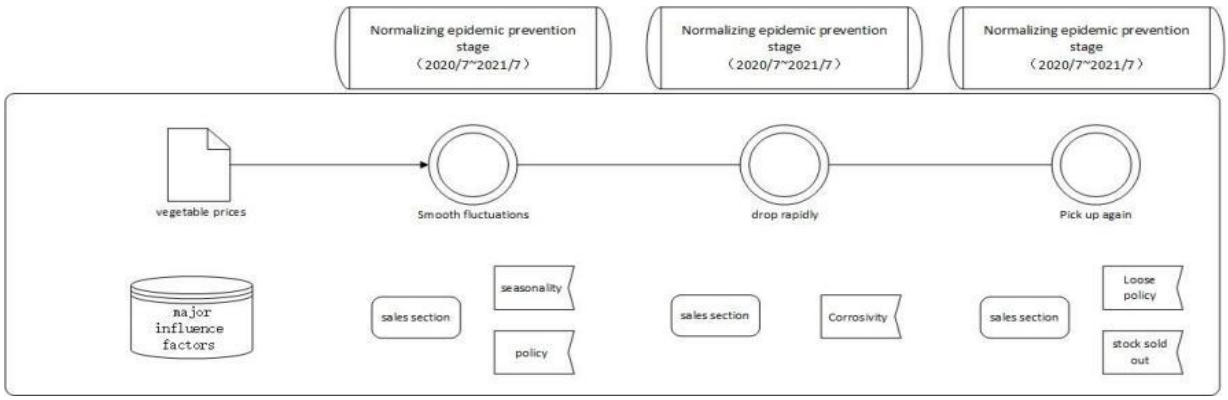

Fig 4. Summary of vegetable price analysis from the perspective of supply chain.

Comparing Figures 3 and 4 reveals the significant impact of the pandemic on the sales of fresh supermarkets. Sales fluctuate within a certain range during normal epidemic prevention due to policy regulations, decline during outbreaks, and rebound as the epidemic is controlled, with a rapid market recovery. Notably, in November 2022, sales stabilized as the retail industry improved and online consumption surged. This demonstrates the varying degrees of impact the pandemic's different stages have on the supply chain and subsequent sales.

2.4. Correlation analysis between categories based on Pearson correlation coefficient

2.4.1. Visualization of line chart

Firstly, the visual presentation of category preprocessing is shown in Figure 5. Since the fluctuation patterns of each line are similar, the use of the Pearson correlation coefficient can be considered to quantify the linear correlation between them [7].

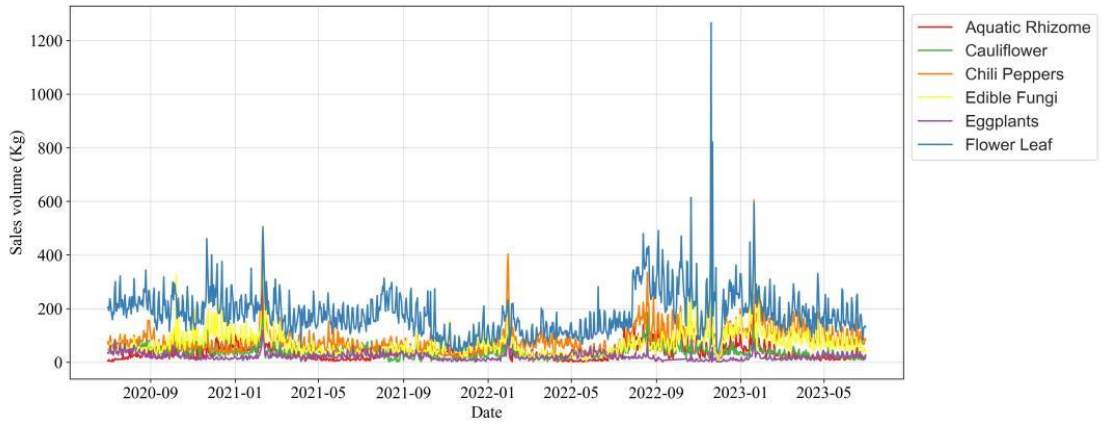

Fig 5. Schematic diagram of daily sales volume of each category over time.

2.4.2. Calculate Pearson correlation coefficient

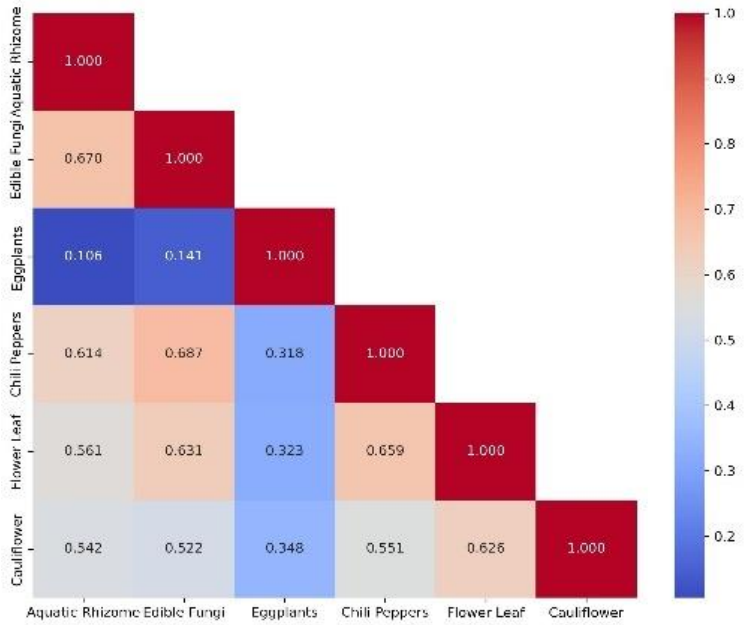

Fig 6. Visualization of Pearson correlation coefficient

Based on Figure 6, it can be observed that the correlation coefficients of the same category towards other categories do not differ significantly. By further observing the line charts of daily sales volume for each category, it is found that certain categories exhibit instability in time series during specific time periods (for example, the flower and leaf category during 2022/11~2023/1). Given that the Pearson correlation coefficient may be difficult to capture such complex non-monotonic relationships, this paper proposes to verify this hypothesis through the ADF test.

2.4.3. ADF stability test

ADF test is a statistical method for detecting the stationarity of time series [8], based on unit root testing. In this study, six time series were analyzed for stationarity using Python's adfuller function, and the ADF test chart is shown in Figure 7. Note that a p-value greater than 0.05 indicates acceptance of the null hypothesis, meaning the series is non-stationary.

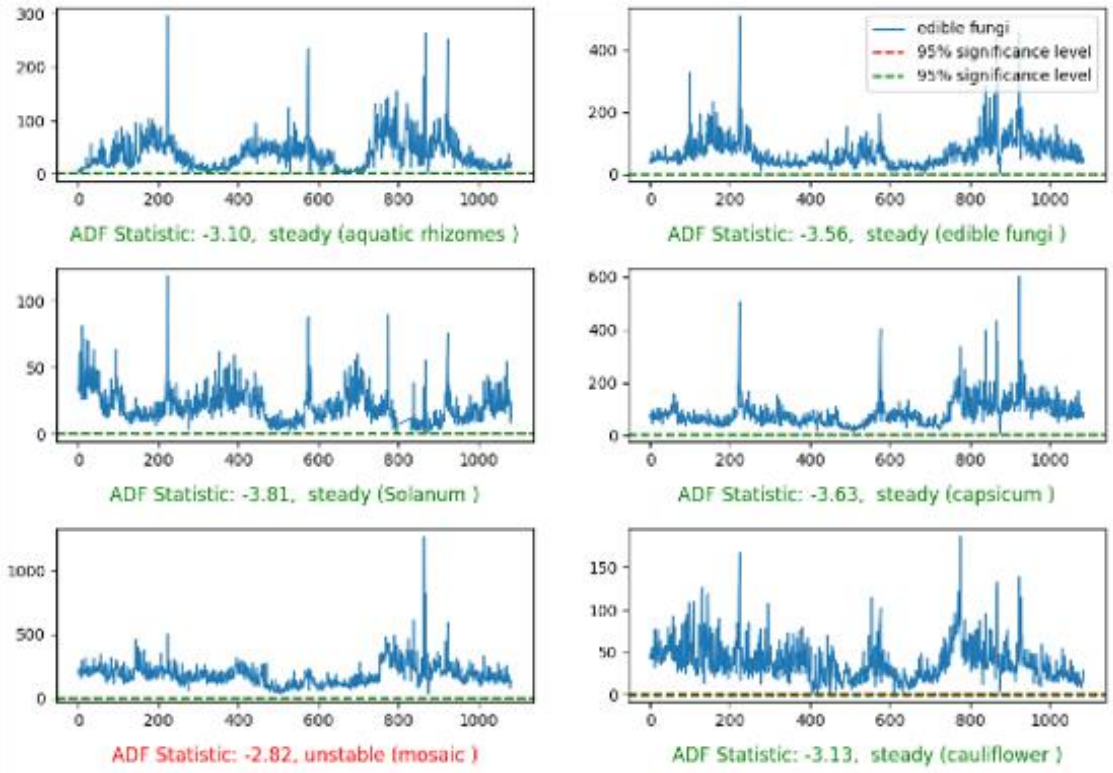

Fig 7. ADF test chart of each time series.

Through in-depth analysis of the ADF test chart, the non-stationarity of the flower and leaf category time series is verified. Therefore, to more accurately analyze the data, this paper adopts a more suitable DPCCA model for re-modeling and analysis.

2.5. Partial correlation analysis between categories based on DPCCA

The analysis process of inter-category partial correlation based on the DPCCA method mainly consists of the following six steps:

(1) For 6 time series $\{x_i^1\},\{x_i^2\},\{x_i^3\},\{x_i^4\},\{x_i^5\},\{x_i^6\}$, each time series can be viewed as a random walk process, so a cumulative sum sequence is constructed for it. Among them, $j=1,2,3\cdots 6, k=1,2,3\cdots N$ (here: $N=1085$).

(2) Divide the entire sequence into several non-overlapping windows, each containing a fixed number of values. For each window, fit the data using a polynomial to capture local trends. Subsequently, define the detrended walk as the difference between the original time series and the fitted local trend sequence.

(3) Therefore, for each time series $\{x_i^j\}$, a residual time series after trend removal is obtained $Y_l^j, l=1,2,3\cdots,(N-s)(s+1)$. Based on the residual time series, the covariance between the two can be further calculated and the covariance matrix can be written, where $j1, j2=1,2,3\cdots,m$.

(4) Calculate the cross-correlation between any two time series, and construct a sparse cross-correlation matrix based on this.

(5) The relationship between $\{x_i^{j_1}\}$ and $\{x_i^{j_2}\}$ can be reflected through the coefficient matrix, but this method cannot exclude the possibility that the two sequences may be simultaneously affected by other time series, which may introduce errors. To solve this problem, a cross-calculation method is adopted to eliminate the potential influence of other time series, and then calculate the inverse matrix of $\rho(s)$.

(6) Define the cross-correlation level between $\{x_i^{j_1}\}$ and $\{x_i^{j_2}\}$, where a subset of cross-correlation coefficients, denoted as $\rho_{DPCCA}(j_1, j_2; s)$, can be used to measure the cross-correlation level between the two time series at a time scale s, excluding the influence of other time series.

$$P_k^j = \sum_{i=1}^{k} \mathrm{x}_i^j \tag{1}$$

$$Y_{(i-1)(s+1)+k-i+1}^j = P_k^j - P_{k,i}^j \tag{2}$$

$$F_{j1,j2}^2(s) = \frac{\sum_{l=1}^{(N-s)(s+1)} Y_l^{j1} Y_l^{j2}}{(N-s)(s-1)}, F^2(s) = \begin{pmatrix} F_{1,1}^2(s) & \cdots & F_{1,m}^2(s) \\ \vdots & \ddots & \vdots \\ F_{m,1}^2(s) & \cdots & F_{m,m}^2(s) \end{pmatrix} \tag{3}$$

$$\rho_{j1,j2}(s) = \frac{F_{j1,j2}^2(s)}{F_{j1,j1}(s) F_{j2,j2}(s)}, \rho(s) = \begin{pmatrix} \rho_{1,1}(s) & \cdots & \rho_{1,m}(s) \\ \vdots & \ddots & \vdots \\ \rho_{m,1}(s) & \cdots & \rho_{m,m}(s) \end{pmatrix} \tag{4}$$

$$C(s) = \rho^{-1}(s) = \begin{pmatrix} C_{1,1}(s) & \cdots & C_{1,m}(s) \\ \vdots & \ddots & \vdots \\ C_{m,1}(s) & \cdots & C_{m,m}(s) \end{pmatrix} \tag{5}$$

$$\rho_{DPCCA}(j_1, j_2; s) = \frac{-C_{j_1 \cdot j_2}(s)}{\sqrt{C_{j_1 \cdot j_1}(s) \cdot C_{j_2 \cdot j_2}(s)}} \tag{6}$$

Based on the above analysis, the final heatmap of the partial cross-correlation coefficient matrix can be generated, as shown in Figure 8 (Note: Due to mathematical properties, the time series' autocorrelation coefficient is set to -1, without affecting its correlation with other sequences.).

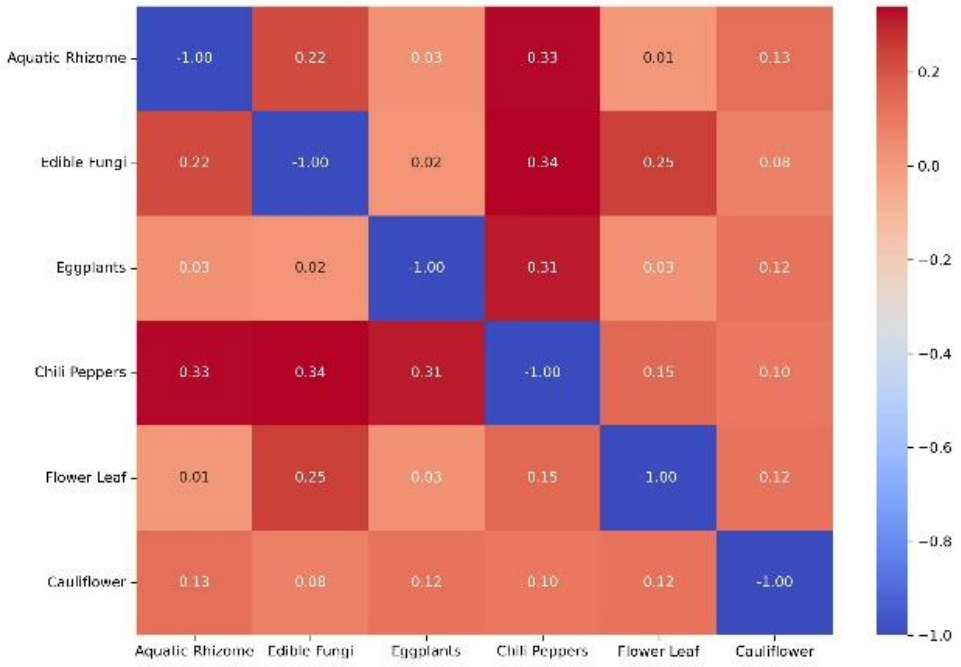

Fig 8. Thermodynamic diagram of cross-correlation coefficient matrix.

Analysis of the DPCCA algorithm indicates that certain categories within the vegetable group [9], including aquatic roots, edible fungi, peppers, eggplants, flowers, and cauliflowers, exhibit positive correlations. The correlation between aquatic roots, edible fungi, and peppers is particularly strong, ranging from 0.2 to 0.3. Conversely, the correlation between aquatic roots and eggplants, as well as between edible fungi and eggplants or cauliflowers, is relatively weak, ranging from 0 to 0.1. These findings offer valuable insights for vegetable cultivation and supply chain management.

3. Analysis of replenishment and pricing strategies for the upcoming week

3.1. Establishment of the model for maximizing returns

The establishment of the model for maximizing returns mainly consists of the following eight steps:

(1) Based on the previous research and design of the profit maximization scheme, an objective function is established. In this function, W_i represents the profit obtained on the i th day in the future; X_i represents the sales volume on the i th day in the future; λ_i and s_i represent the pricing and purchase price on that day, respectively.

(2) Given that some goods may experience losses during sales, there is a certain mathematical relationship between sales volume X_i, purchase volume M_i, and loss rate ∂. In order to accurately describe this relationship, this article studies and establishes corresponding expressions.

(3) Since the loss rate has remained at an average level in the near future, it is replaced by the average loss rate $\bar{\partial}$.

(4) Given that there is a certain functional relationship between pricing and daily sales, a corresponding mathematical expression can be constructed to describe this relationship. Where $g_j(\lambda_i)$ represents the fitting function of the sales volume of category j on the pricing on day i.

(5) Future pricing s_i is predicted through time series, and since the error is negligible, the predicted value $\tilde{s}_i$ is used instead.

(6) Based on the aforementioned analysis steps, this paper derives a specific mathematical model for achieving profit maximization.

(7) Based on sales volume and loss, the purchase quantity needs to be dynamically adjusted to ensure a balance between supply and demand, so it is described by a function.

(8) To ensure that profits comply with the regulations of the price management department, this paper sets a clear range limit for them.

$$\max W_i = \max(X_i \cdot \lambda_i - M_i \cdot s_i), i = 1,2,3\cdots 7 \tag{7}$$

$$X_i = M_i \cdot (1-\partial_i) \tag{8}$$

$$\partial = \bar{\partial} \tag{9}$$

$$X_i = g_j(\lambda_i) \tag{10}$$

$$s_i = \tilde{s}_i \tag{11}$$

$$\max W_i = \max(\lambda_i \cdot g_j(\lambda_i) - \frac{g_j(\lambda_i)}{1-\partial(\lambda_i)} \cdot s_i), i = 1,2,3\cdots 7,\ j = 1,2,3\cdots 6 \tag{12}$$

$$M_i = \frac{g(\lambda_i)}{(1-\partial(\lambda_i))} \tag{13}$$

$$0.25 \le \frac{\lambda_i - s_i}{\lambda_i} \le 0.75 \tag{14}$$

3.2. Analysis of future pricing based on ARIMA

3.2.1. Establishment of ARIMA model

ARIMA is a method used for linear time series forecasting [10]. Based on the previous analysis, pricing strategies belong to non-stationary time series with non-specific trends. Therefore, the introduction of probabilistic analysis can be used to seek a stochastic model to improve the ARIMA model in order to make a more accurate fit to the sequence.

In the ARIMA (p, d, and q) model, AR (p) predicts future changes via historical data, MA (q) denoises with sliding smoothing, and d represents differencing order. ARIMA's model.

$$y_t = \mu + \varepsilon_t + \sum_{i=1}^{p} \gamma_i y_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{1-i} \tag{15}$$

Among them, y_t represents the sequence value at time t , ε_t represents the residual value at time t , μ represents the constant term, the autocorrelation coefficient is represented by γ_i , and p and q represent the order of AR and MA , respectively. θ_i Represents the moving average coefficient. $\sum_{i=1}^{p} \gamma_i y_{t-i}$ Is the autoregressive part of ARIMA, which can capture and reflect the temporal nature of the sequence? The rest are disturbance values, which are often related to external factors that remove the temporal pattern.

Based on the stationarity test of the time series for each category, the results show that the water-rooted stem vegetables are non-stationary, while edible fungi, eggplant, pepper, leaves and flowers vegetables exhibit stable characteristics.

3.2.2. Determination of model parameters

To determine the relevant parameters of the model, the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are typically used to construct model parameters, continuously optimizing the model to minimize both values.

For stationary sequences, initially set d=0, then select parameter combinations of p=0~5 and q=0~5 to find the optimal parameters and test the autocorrelation of residuals. If the test passes, further differencing is unnecessary. For non-stationary sequences, all parameter combinations of p=0~5, d=0~2, and q=0~5 are chosen through brute-force calculation to automatically select the best model parameters under different combinations of p, d, and q. The final model parameters are shown in Table1:

Table 1. Parameters of ARIMA model for each category.

Category	Parameters (p, d, q)	(AIC,BIC)
Aquatic Rhizome	ARIMA(2,1,3)	(3610.336,3640.266)
Edible Fungi	ARIMA(1,0,2)	(3053.209,3078.156)
Eggplants	ARIMA(1,0,2)	(1675.898,1700.681)
Chili Peppers	ARIMA(1,0,1)	(2304.080,2324.038)
Flower Leaf Veg	ARIMA(1,0,1)	(1139.271,1159.228)
Cauliflower Veg	ARIMA(1,0,0)	(1524.544,1539.510)

3.2.3. Model validity verification

After selecting model parameters for each category, it is necessary to verify their suitability. The autocorrelation of residuals is an important criterion for assessing the applicability of the ARIMA model. The closer the residuals are to a normal distribution and uncorrelated, the more accurately the model captures useful information from the original time series. Residual signal distribution and ACF plots were created for each category, resulting in Figure 9 (Due to the similar distribution of the remaining images to the ones already presented, they are omitted here to avoid repetition.):

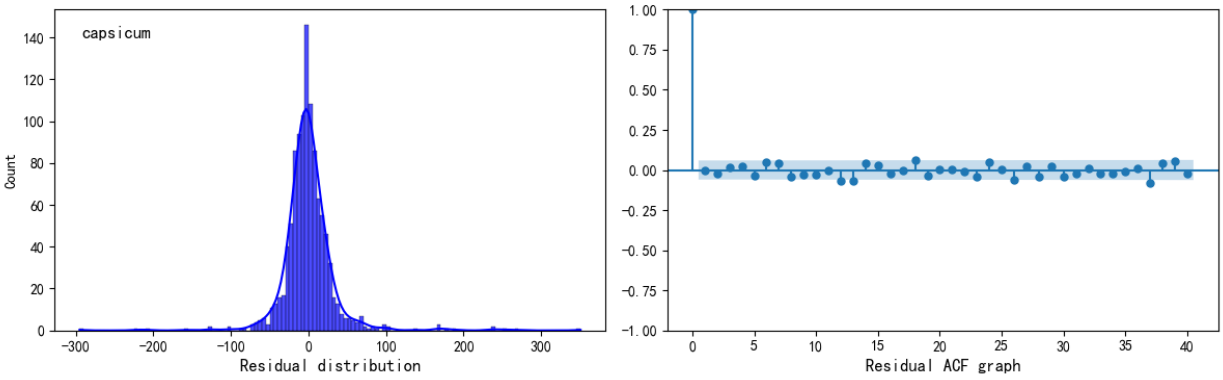

Fig 9. Visualization of Spearman correlation coefficient.

As can be seen from Figure 9, the residual distribution plots for each category under the model are all approximately normally distributed, and the ACF plots of the residuals almost entirely fall within the horizontal line area, indicating that the validity of the model has been verified.

3.2.4. Future pricing prediction

After selecting the optimal model parameters for each category, the adjusted ARIMA model was used to predict future purchase prices. The prediction results are shown in Table 2.

Table 2. Prediction of the purchase price of various dishes in the coming week.

Next week	Aquatic Rhizome	Edible Fungi	Eggplants	Chili Peppers	Flower Leaf	Cauliflower
2023/7/1	12.597	7.869	4.963	4.699	3.151	7.822
2023/7/2	12.610	7.596	4.973	4.728	3.168	7.745
2023/7/3	12.564	7.600	4.988	4.757	3.185	7.671
2023/7/4	12.627	7.603	5.003	4.785	3.202	7.601
2023/7/5	12.564	7.607	5.017	4.813	3.218	7.535
2023/7/6	12.613	7.611	5.030	4.840	3.233	7.472
2023/7/7	12.587	7.614	5.044	4.867	3.248	7.412

4. Conclusions

This paper delves into the pricing and replenishment strategies for fresh supermarkets to maximize revenue through a study based on time series and supply chain perspectives. The content primarily focuses on the sales distribution patterns of vegetable categories and individual products. Through visual analysis, seasonal sales patterns are revealed, and the ADF test and DPCCA model are employed to explore the correlation between categories. The study further utilizes sales data from the past month, predicts pricing using the ARIMA model, and combines it with nonlinear programming methods to formulate optimal pricing and daily replenishment strategies for the next seven days. The conclusion demonstrates that by integrating time series analysis and supply chain management, fresh supermarkets can more accurately predict sales trends and develop effective replenishment and pricing strategies accordingly, thereby maximizing revenue. In future research, AI and big data can boost prediction accuracy in supermarket revenue modeling. Incorporating external factors like climate and marketing will lead to more flexible strategies, fostering sustainable supermarket growth.

References

[1] Zheng Yan, Huang Xing, Xiao Yujie. Research on Commodity Demand Forecasting Model Based on Time Series [J]. Journal of Chongqing University of Technology (Natural Science), 2019, 33(09): 217-222.

[2] Cui Yunhao. Research on Sales Forecast of Fresh Vegetables Based on CNN-PSO-LSTM Combined Model [D]. Anhui Agricultural University, 2023.

[3] Li Zhonglin, Li Tianyu, Liu Zeyu, et al. Catering Market Demand Forecasting Model Based on Time Series Analysis [J]. Scientific and Technological Innovation, 2023(10): 89-92.

[4] Chen Yixuan. Optimization Strategy of Fresh Agricultural Products Supply Chain [J]. Business Culture, 2021(15): 28-29.

[5] Li Jiaojiao, He Lili, Zheng Junhong. Inventory Cost Control Model of Fresh Agricultural Products Retailers Based on DDQN [J]. Intelligent Computer and Applications, 2023, 13(10): 60-64+72.

[6] Luo Qin. Comparative Analysis of Vegetable Retail Price Fluctuations in Chongqing and Nationwide [J]. Northern Horticulture, 2024(02): 139-146.

[7] She Zihang, Xu Jiahua, Yao Zhiyu, et al. Online Shopping Big Data Analysis Based on Pearson Correlation Coefficient - Taking the Transaction Records of Bairunju Flagship Store on Tmall as an Example [J]. Journal of Hanshan Normal University, 2020, 41(03): 16-22.

[8] Yang Yanjun, Tang Di. Real-time Early Warning Research on Price Bubbles in China's Non-ferrous Metals Futures Market - Analysis Based on the Supremum ADF Test Method [J]. Price Theory and Practice, 2022, (12): 114-117+202.

[9] Cao Jianing. Research and Application of Partial Cross-Correlation for Non-stationary Time Series [D]. Beijing Jiaotong University, 2022.

[10] Li Lingling, Xin Hao. Application of Time Series Model ARIMA in Data Analysis [J]. Fujian Computer, 2024, 40(04): 25-29.