Automatic Pricing and Replenishment Decisions for Vegetable Commodities

Xinyu Li, Linsong Li*

Mathematics Department, Yanbian University, Yanji 133002, China

Abstract: The various categories of vegetables show different sales trends with the changes of specific times and seasons. Reasonable arrangements and the formulation of pricing strategies, as well as the timely and effective replenishment of individual items, can increase the profits of supermarkets. Firstly, this article uses the 3σ principle to preprocess and visualize the sales data of six major vegetable categories, including eggplant, cauliflower, pepper, aquatic rhizome, and leafy and floral vegetables, and their corresponding individual vegetable items. The AMRIA time series model is established by using the statistical analysis software SPSS 26.0 to analyze the distribution of the sales volume of vegetable categories and individual items over time. It is found that the overall sales volume of individual items fluctuates over a long period of time, and leafy and floral vegetables, cauliflower, and aquatic rhizome vegetables show seasonal trends. Determine the correlation between various categories and individual products. through the Pearson correlation coefficient. A neural network prediction model is established and solved by using MATLAB software to predict the daily sales volume of each category in the next week. Subsequently, an optimization model is established to determine the total daily replenishment quantity and pricing strategy for vegetable categories in the next week, and a judgment matrix is constructed to determine the relevant data in different aspects required for supermarkets to formulate replenishment and pricing strategies.

Keywords: Correlation Analysis; Spearman Correlation Coefficient; Neural Network Prediction Model; Optimization Model

DOI: 10.57237/j.wjms.2025.01.003

1 Introduction

In fresh food supermarkets, the shelf life of general vegetable products is relatively short, and their appearance tends to deteriorate as the sales time increases. Most varieties cannot be sold the next day if they are not sold on the same day. Therefore, supermarkets usually replenish their stock every day based on the historical sales and demand of each product. [1]

Since there are numerous vegetable varieties sold in supermarkets and their places of origin vary, and the purchase and trading time of vegetables is usually between 3:00 and 4:00 in the early morning, merchants have to make replenishment decisions for each vegetable category on the same day without knowing exactly the specific individual items and purchase prices. Vegetables are generally priced using the "cost-plus pricing" method, and supermarkets usually offer discounts on goods with transportation damage or deteriorated appearance. [2]

Reliable market demand analysis is particularly important for replenishment decisions and pricing decisions. This article uses the ARIMA model to find out the distribution law of the sales volume of categories and individual items over time, and uses the Pearson correlation coefficient to identify the correlations among various categories and individual items [3]. A neural network predic-

*Corresponding author: Linsong Li, 3284434795@qq.com

Received: 20 January 2025; Accepted: 6 May 2025; Published Online: 30 June 2025 http://www.wjoms.com

tion model is established to predict the sales data for the next week, and the total daily replenishment quantity and pricing strategy for each vegetable category in the next week are provided to maximize the profits of supermarkets. Moreover, a judgment matrix model is established to solve the problem of what relevant data supermarkets still need to collect in order to better formulate replenishment and pricing decisions for vegetable products and how these data can help solve the above-mentioned problems.

2 Preprocessing of Commodity Sales Data

(1) Judging outliers based on the 3σ principle

There are some extreme values for the unit sales price in the table. In order to determine these outliers more quickly and accurately, we use the 3σ principle to judge the outliers. It can be seen from the figure that when the data obeys the normal distribution, 99.7% of the data falls within three standard deviations. Therefore, only 0.3% of the data falls outside it. So the data falling outside three standard deviations is in the minority and can be regarded as a low-probability event, and the data within this range can be regarded as outliers.

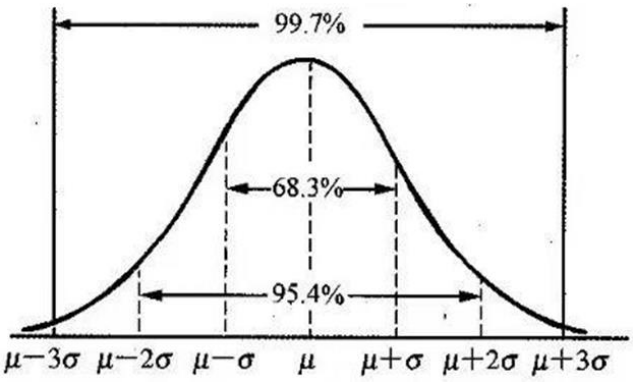

Figure 1 Judging outliers based on the 3σ principle

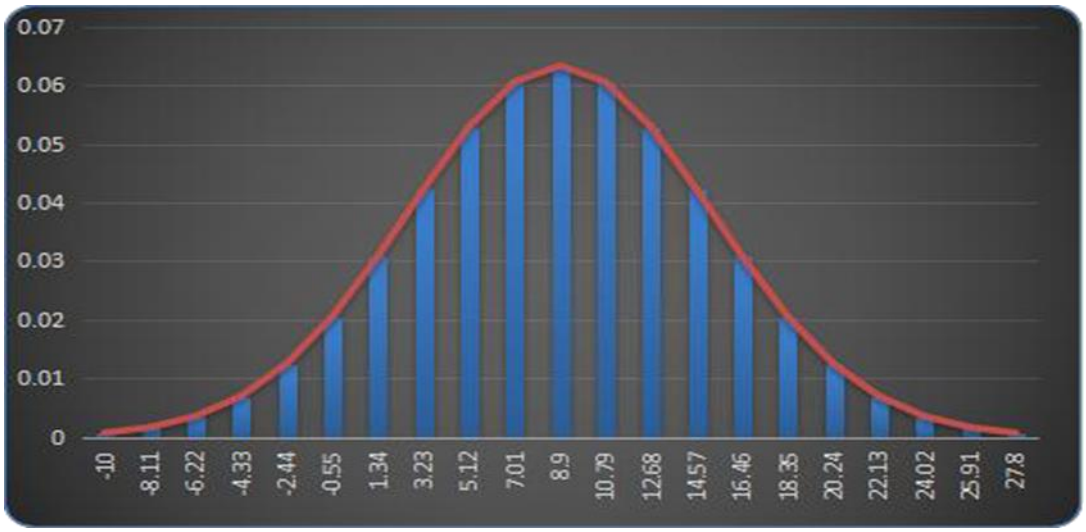

Figure 2 Normal distribution graph

We can use an Excel spreadsheet to calculate the mean and standard deviation of the unit sales price [4]. The obtained mean is 8.9 and the standard deviation is 6.3.

Secondly, we can calculate the probability density corresponding to each unit sales price. Taking the unit sales price as the abscissa and the probability density as the ordinate, we can draw a normal distribution graph as shown in the figure. It can be seen from the figure that prices above 27.8 are outliers.

We also make a scatter plot of single product code - unit sales price, as shown in the figure. There is a sharp upward trend in the red circles, so we can determine that there are outliers.

(2) Outlier processing

After we conduct outlier determination on the data set, we screen out the outliers and perform elimination processing after outlier determination on the collected samples. Eventually, we can determine the outliers through logical analysis and the 3σ principle. We can use an Excel spreadsheet to label the obtained outliers. We assume that the outliers are labeled as "a", and search for data "a" by means of screening. For the data where the outliers are located, we conduct elimination processing. Therefore, it is determined that among the 1,095 pieces of data, there

are a total of 123 out-of-range data and 25 data with a value of 0, totaling 158 outliers. After all of them are removed, there are 937 days' worth of normal daily sales data left in total.

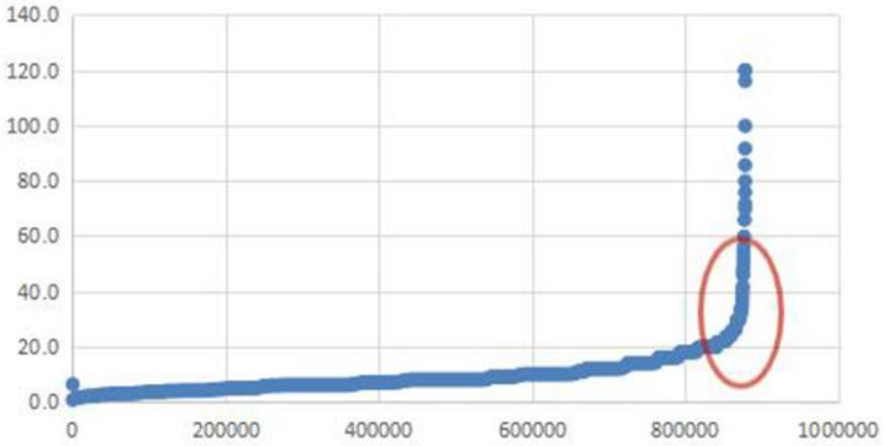

Figure 3 A scatter plot of single product code

3 Establishment and Solution of the ARIMA Model and the Correlation Coefficient Model

(1) Use the ARIMA model to find out the distribution law of the sales volume of categories and single products over time.

For the distribution law of the sales volume of vegetable categories and single products over time, use the ARIMA model for time series analysis. The basic idea of the ARIMA model is to use the historical information of the data itself to predict the future. It mainly consists of three parts, namely the autoregressive model (AR), the differencing process (I), and the moving average model (MA) [5].

The original graph of the distribution of single products over time in the recent three years is shown in Figure 4.

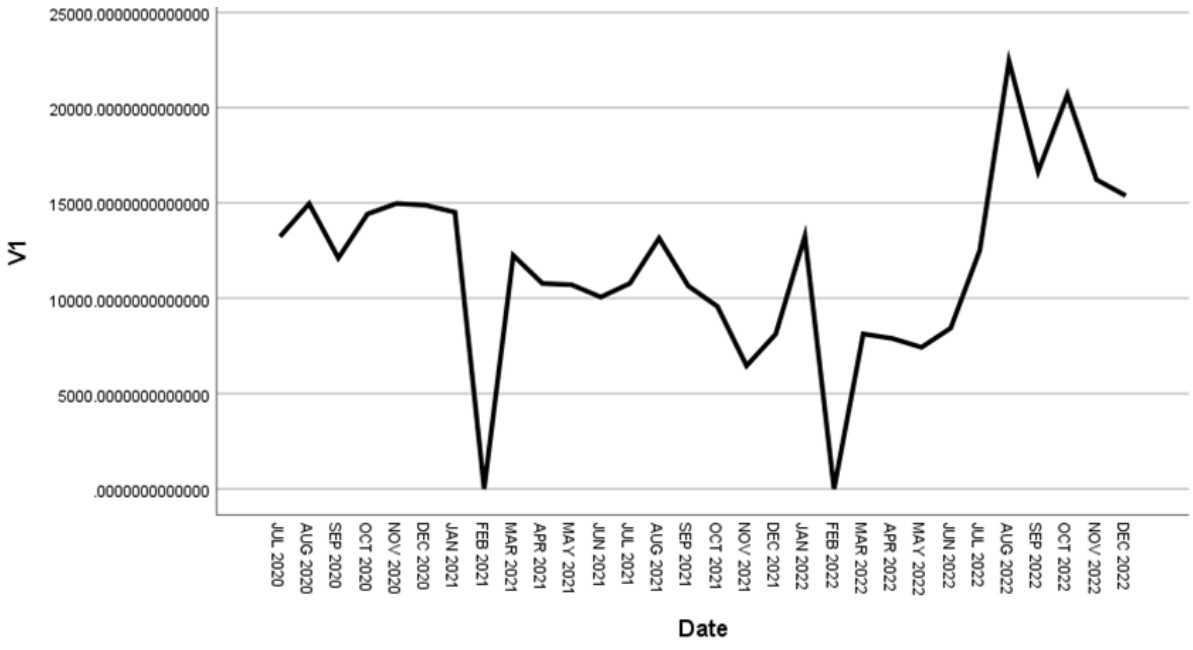

Figure 4 Original Graph of the Distribution of Single Products over the Past Three Years

It can be seen that there are certain seasonal fluctuations. Conduct the ADF test on it and determine that the order of differencing is 2. The obtained stationary data is shown in Figure 5.

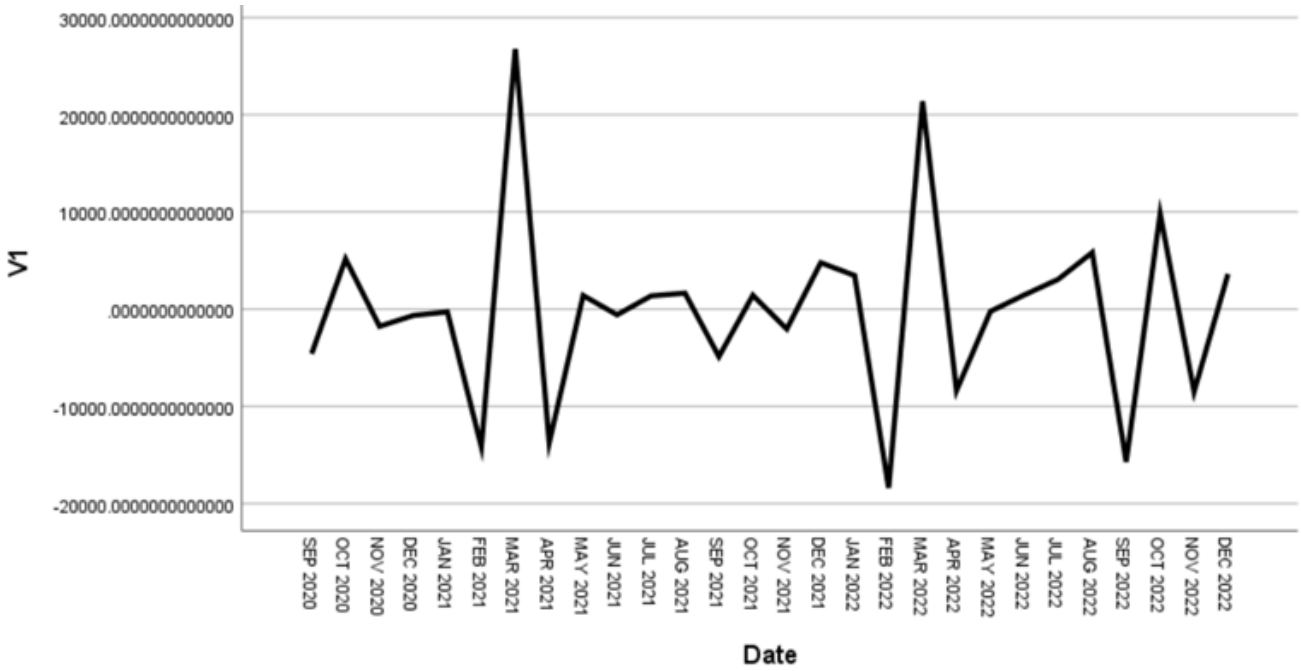

Figure 5 Stationary Data Graph Obtained with the Differencing Order of 2

Use SPSS to determine that the parameters p and q are 1 and 1 respectively.

The distribution of the sales volume of single product data over time is shown in Figure 6.

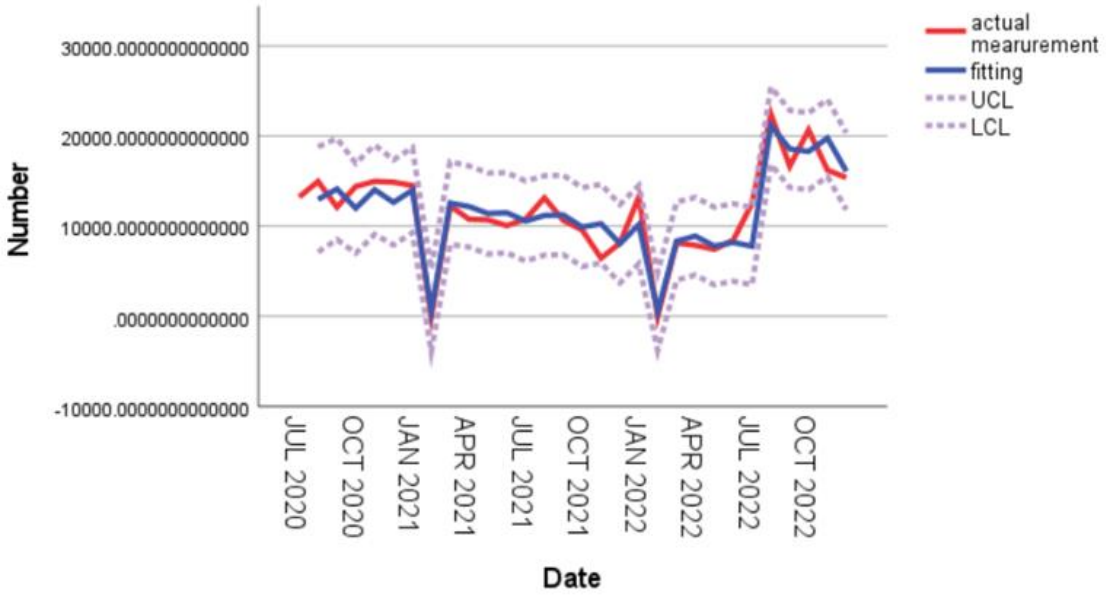

Figure 6 Time Series of Single Product Sales

There are peaks and troughs in the sales volume of single products in the figure, and it can be seen that the sales volume of single products presents a certain seasonal distribution.

In the same way, the time distribution diagram of the sales volume of each vegetable category is shown in Figure 7.

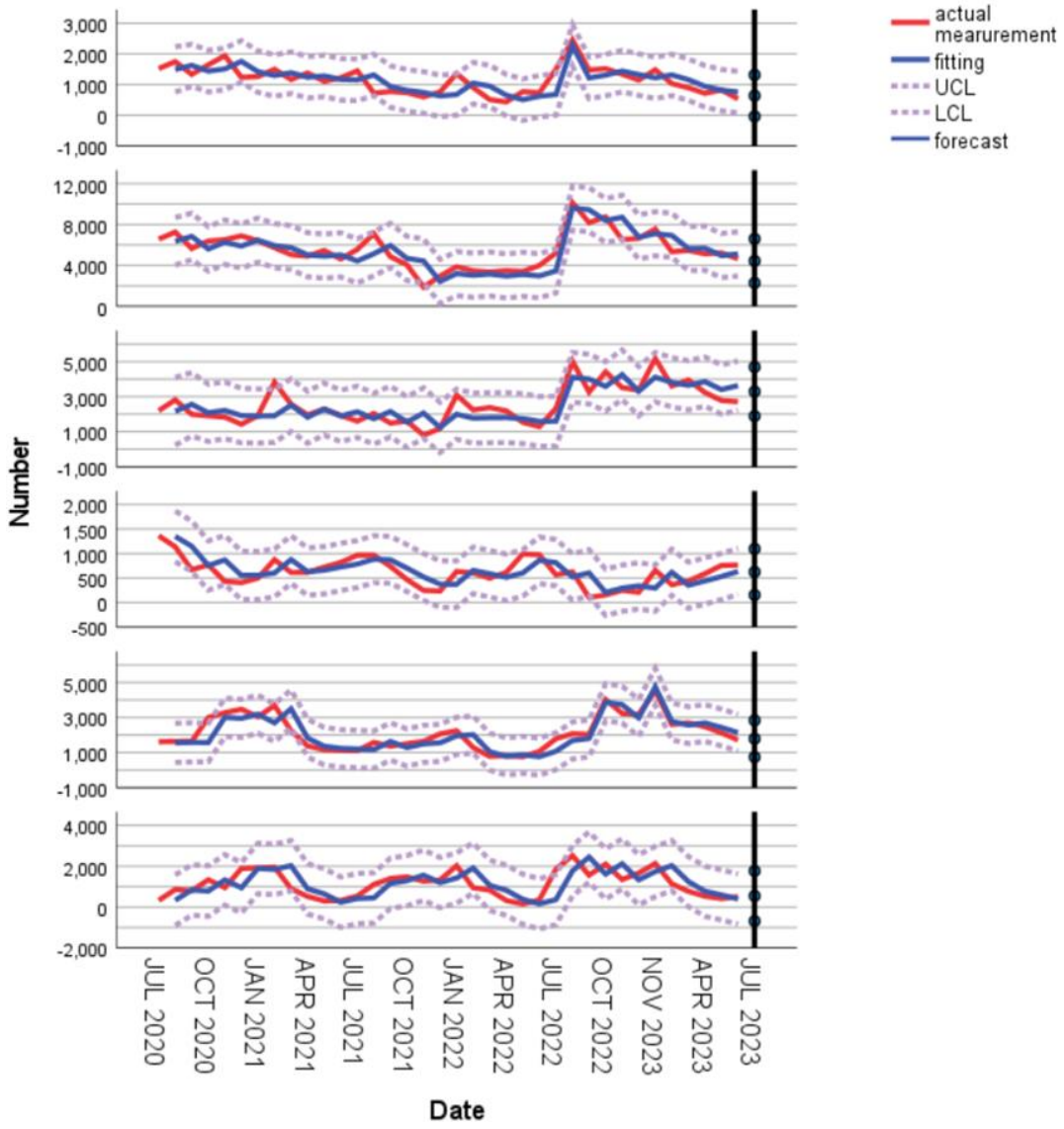

Figure 7 Time Distribution Diagram of the Sales Volume of Each Vegetable Category

It can be seen that vegetable categories also exhibit certain seasonal fluctuations. Meanwhile, it can also be found that the monthly sales volume proportions of cauliflower, leafy and aquatic rhizome categories from July 2022 to June 2021 are very similar, and the sales volumes of all categories declined between October 2022 and December 2022, and the decline range was greater than that in other years.

(2) Use the Pearson correlation coefficient to find the correlations among various categories and single products.

The Pearson correlation coefficient is used to measure the degree of correlation between two variables. It can rank the correlations between variables and analyze the correlations between two continuous variables [6]. Since the sales volumes of various vegetable categories are continuous variables, the Pearson correlation coefficient is adopted in this modeling to analyze the correlation relationships of the sales volumes of various vegetable categories [7]. As shown below.

$$\rho(x,y)=\frac{\sum_{i=0}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=0}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=0}^{n}(Y_i-\bar{Y})^2}}$$

Among them, $X_i$and $Y_i$respectively represent the sample sequence values of the monthly sales volumes of two vegetable categories, and $\bar{X},\bar{Y}$ represent the sample means of the sample sequences.

Use Matlab to draw the heat map of the correlation coefficients of each category as shown in Figure 9.

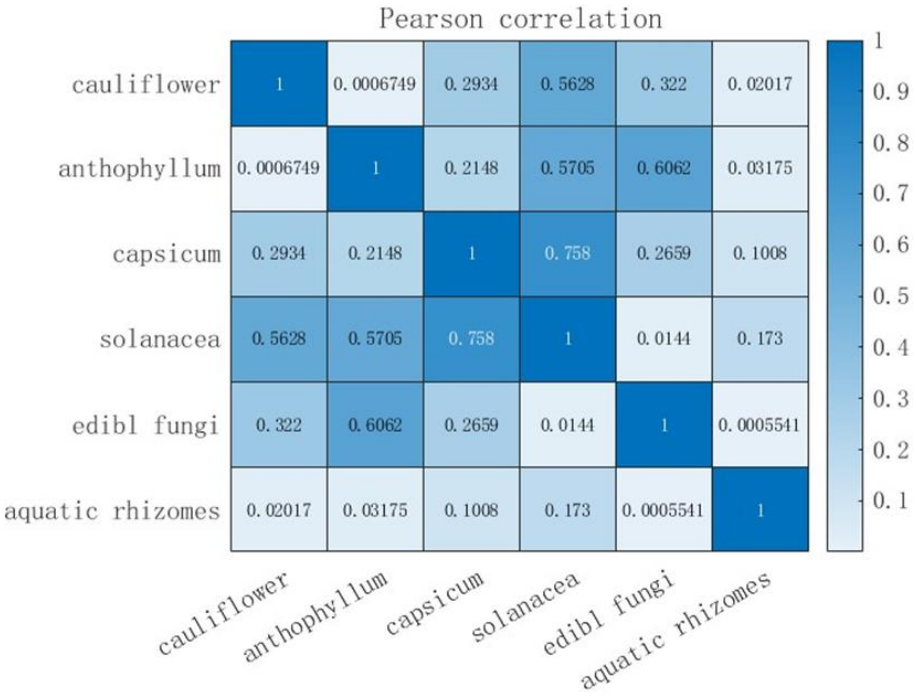

Figure 8 The heat map of the correlation coefficients of each category

We use the Spearman correlation coefficient (define X and Y as two sets of data, where di is the rank difference between Xi and Yi) to obtain the correlation coefficients among various single products.

The heat map is drawn as follows:

单品编码 / 年月	102900005115168	102900005115199	102900005115250	102900005115625	102900005115748	102900005115762	102900005115779	10
1	56.624	0.000	202.540	0.000	19.904	17.538	807.989	
2	5.001	3.692	740.681	0.000	160.665	71.991	868.321	
3	109.751	169.838	461.956	0.000	503.454	474.566	1011.025	
4	0.000	147.801	436.653	0.000	0.000	1077.352	913.197	
5	344.027	5.978	396.292	0.000	0.000	1008.428	1130.132	
6	62.346	0.000	310.012	0.000	0.000	888.322	1350.552	
7	0.000	0.000	0.273	0.000	0.000	635.247	2985.501	
8	0.000	0.000	0.000	0.000	0.000	519.714	2144.176	
9	0.000	0.000	98.709	3.381	0.000	302.203	1779.850	
10	13.678	0.000	0.576	70.370	0.000	101.798	1576.135	
11	13.126	6.176	164.058	44.004	22.950	5.499	758.014	
12	295.284	0.000	1.965	3.265	11.703	0.000	590.530	

12 rows × 246 columns

Figure 9 The correlation coefficients among various single products

Draw a heat map as follows

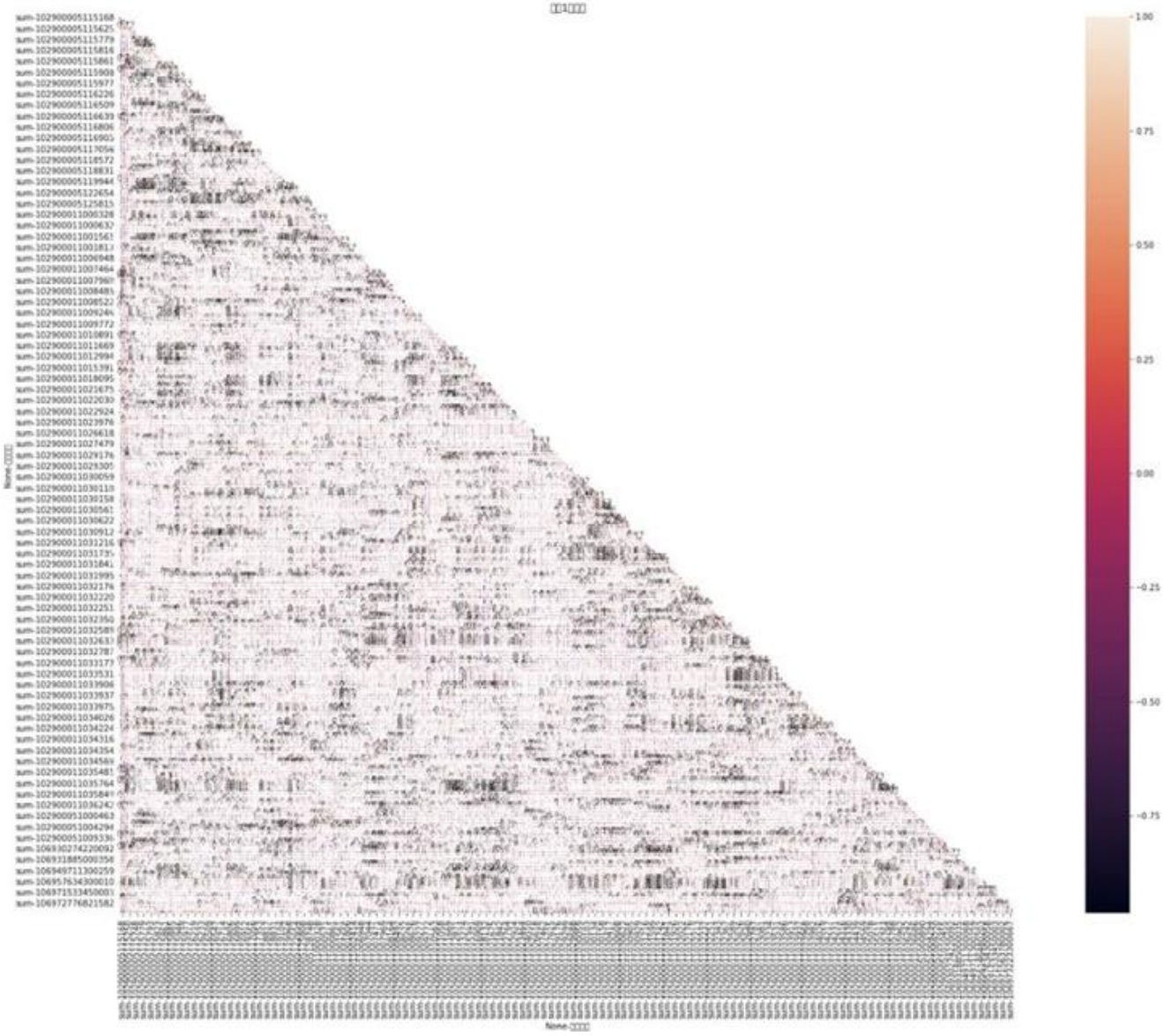

Figure 10 A heat map

Table 1 Sales data

Item No. 1	Item No. 2	correlation coefficient
102900011033562	102900011033531	1.0
102900011033586	102900011033562	1.0
102900011033586	102900011033531	1.0

From the table in the figure above and the data in the attachments provided in the question, it can be known that these three groups of single products have the same purchase time for customers. We can learn from it that these single products have a collocation relationship. Through this collocation relationship, the correlation between pairs of single products can be better strengthened, making the sales volume of single products more stable.

4 Establishment and Solution of the BP Neural Network Model

Based on the historical data in the given attachments, we use neural network prediction to forecast the daily sales quantity of each category within the next week. The BP neural network is a kind of artificial neural network, and its main

function is to conduct distributed information processing [8].

Its core steps are as follows:

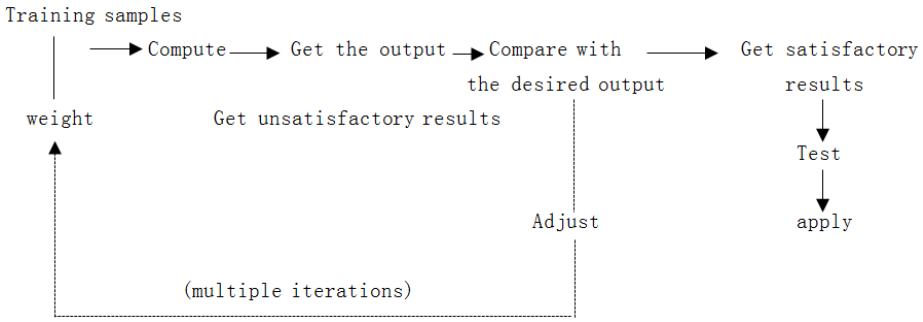

Figure 11 Core steps

Step length =$\eta\frac{df(x)}{dx}$

According to this formula, the value of the next point is obtained as follows:$x^{(k+1)}=x^k-\eta\frac{df(x)}{dx}$

Before derivation for:

$$Loss=\frac{1}{2}\sum_{i=1}^{n}(y_i-y_i^*)$$

This is the error term.

The main content of its backpropagation is as follows:

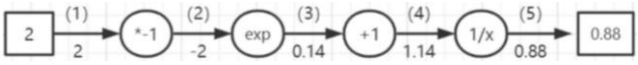

Figure 12 Backpropagation

By the operation of the neural network prediction step, we can derive information about

$$y(t)=f(y(t-n), y(t-n+1),\ldots, y(t-1))$$

BP Comparison of the predicted and actual values of the neural network training set [9].

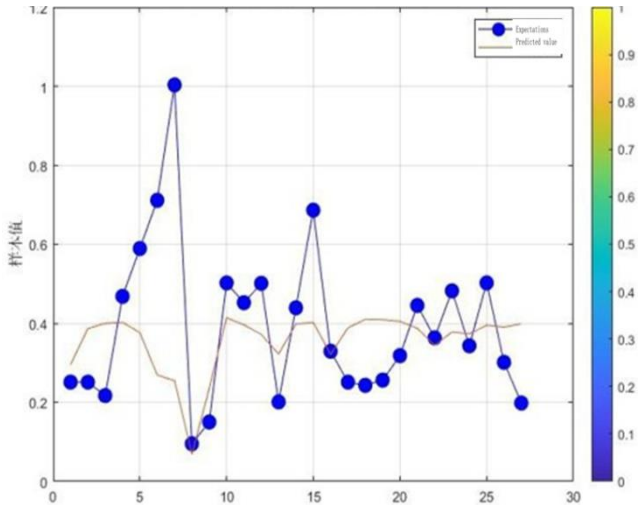

Figure 13 BP Comparison of the predicted and actual values of the neural network training

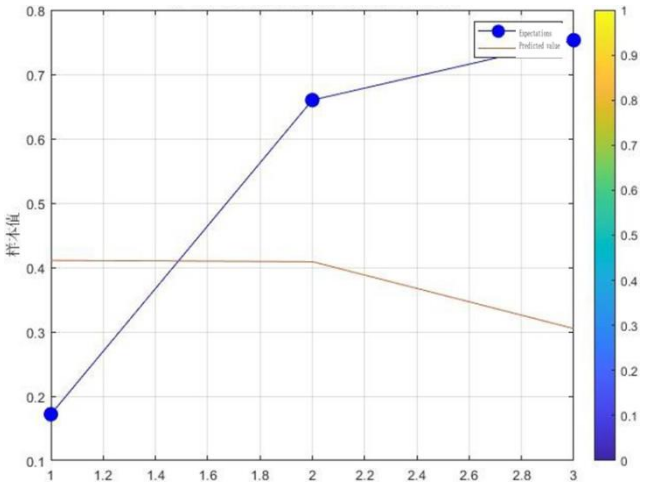

Figure 14 Neural network training

Through the operation of the neural network prediction steps, it is possible to summarize the data based on the previous question and then predict the daily sales quantity of each category in the future.

The above figure clearly shows the relationship between the predicted values and the actual values, which can help to better determine the pricing strategy and the total replenishment quantity.

5 Establishment and Solution of the Model for Question Three

The objective function is determined as maximizing ∑ (predicted sales volume × (sales price - wholesale price)) × (1 - loss rate). The following table shows the single products, unit prices and sales volumes for one week in the latter half of June.

Table 2 The single products, unit prices and sales volumes for one week in the latter half of June

	The name of the item	Sales volume (kg)	Classification name	Sales unit price (RMB/kg)
0	Colorful Pepper (2)	7.302	Chili peppers	23.6
8	Colorful Pepper (2)	7.302	Chili peppers	33.6
12	Colorful Pepper (2)	7.302	Chili peppers	21.6
13	Colorful Pepper (2)	7.302	Chili peppers	13.0
14	Colorful Pepper (2)	7.302	Chili peppers	18.0

Based on this table, select representative single products with high profits to conduct predictions and formulate pricing strategies and replenishment quantities.

Table 3 Pricing strategies and replenishment quantities

The name of the item	Expectedprice	Availability
Tremella (flowers)	17.0	2.5
oyster mushroom	31.0	2.5
Chinese cabbage	27.0	2.5
Pure Lotus Root (1)	23.0	2.5
water caltrop	29.7	2.5

6 Establishment and Solution of the Judgment Matrix Model

1. Index data such as special demands during holidays and seasonal demands can be added. In this way, the number of decision variables and constraint conditions for Question Two and Question Three can be increased. Then, by comparing with the pricing strategies given in Question Two and Question Three, we can observe the changes in profits, so as to judge whether the previously formulated pricing strategies are good or bad. This is conducive to further modifications and optimizations, and further strengthens the rigor and feasibility of the strategies.
2. Relevant data of the supply chain and marketers can be added. Use these data to build an evaluation model to evaluate the replenishment policies and pricing strategies given in Question Two and Question Three. A deep understanding of the supply chain and marketers helps to understand the cost of corresponding commodities, thereby increasing the rationality of commodity pricing and maximizing profits. Having a better grasp of the replenishment channels, transportation methods and replenishment time, optimizing the supply chain and reducing waste will be of great help in formulating more reasonable, efficient and safe replenishment policies, ensuring the quality of commodities and their timely supply.
3. Data on competitors' prices and customer feedback results can be added. A full understanding of competitors' price data enables supermarkets and hypermarkets to better understand the market and adjust and optimize their own pricing in a timely manner, which is helpful for better benign market competition. In addition, a full grasp of customer feedback results can help identify shortcomings and deficiencies, which has a positive impact on the further optimization of product sales and the improvement of customer satisfaction [10].

(1) Construct the judgment matrix

The hierarchical structure reflects the relationships among factors, but the proportions of each criterion in the criterion layer in the target measurement are not necessarily the same [11]. In the minds of decision-makers, they each account for a certain proportion [12].

Suppose that we now want to compare the influence magnitudes of m variables X ={x_1,…, x_m} on a certain factor T. According to the suggestions of Saaty and others, we adopt the method of pairwise comparison of variables to establish a pairwise comparison matrix to provide reliable data. That is, each time we take two variables x_i and x_j, and use a_{ij} to represent the ratio of the influence magnitudes of x_i and $x_j$on T. All the comparison results are represented by the matrix A = $(a_{ij})_{nn}$, and A is called the pairwise comparison judgment matrix (referred to as the judgment matrix for short) between T - X. It is easy to see that if the ratio of the influence of $x_i$and $x_j$on T is a_{ij}, [13] then the ratio of the influence of x_j and x_i on T should be $a_{ji}=\frac{1}{a_{ij}}$. [14]

Definition 1: If the matrix A = $(a_{ij})_{nn}$ satisfies

i. $a_{ij}>0$,

ii.$a_{ji}=\frac{1}{a_{ij}}(i,\ j\ =1,2,\Lambda,\ m)$

Then it is called a positive reciprocal matrix (it is easy to see that $a_i=1,\ i=1,\ldots,\ m$).

Regarding how to determine the value of a_{ij}, Saaty and others suggest citing the numbers 1 to 9 and their

reciprocals as the scale.

Table 4 The meaning of the scale

scale	meaning
1	Indicates that the two factors are of equal importance
3	It means that the former is slightly more important than the latter compared to the two factors
5	It means that the former is significantly more important than the latter compared to the two factors
	It means that the former is more strongly important than the latter compared to the two factors
7	It means that the former is more important than the latter compared to the two factors
9	Represents the median value of the above adjacent judgments
2, 4, 6, 8	If the ratio of the importance of factor i to factor j is aij, then factor j
reciprocal	The ratio to the importance of factor i is: aji =1/ aij

The meaning of the scale in Table 4.

From a psychological perspective, having too many gradations will exceed people's judgment ability. It not only increases the difficulty of making judgments but also tends to lead to the provision of false data. Saaty and others also used experimental methods to compare the accuracy of people's judgment results under various different scales. The experimental results also show that it is most appropriate to adopt the 1-9 scale.

Judgment matrices of the criterion layer in the table:
B1: Supply chain data
B2: Competitors' pricing data
B3: Inventory data
B4: Vegetable purchasing power in the market
B5: Customer feedback results
B6: Purchase intention of the target objects

Table 5 Judgment matrices of the criterion layer in the table

A	B1	B2	B3	B4	B5	B6
B1	1	3	1	2	1	1/5
B2	1/3	1	1/3	1/5	1/4	1/3
B3	1	3	1	1	1/2	1/2
B4	1/2	5	1	1	3	3
B5	1	4	2	1/3	1	1
B6	5	3	2	1/3	1	1

Table 6 The meaning of the scale

criterion	Supply chain data	Competitor pricing data	Inventory data	Inventory data	Customer feedback knots	Target Object of willingness to buy
Normative hierarchy value	0.18	0.15	0.16	0.17	0.17	0.18

Positive indicator: $x_{ij} = \frac{x_{ij}-min\{x_{1j},\ldots,x_{nj}\}}{max\{x_{1j},\ldots,x_{nj}\}-min\{x_{1j},\ldots,x_{nj}\}}$

Reverse indicator: $x_{ij} = \frac{max\{x_{1j},\ldots,x_{nj}\}-x_{ij}}{max\{x_{1j},\ldots,x_{nj}\}-min\{x_{1j},\ldots,x_{nj}\}}$

Calculate the proportion of the i-th sample value under the j-th indicator to this indicator [15]: $p_{ij} = \frac{x_{ij}}{\sum_{i=1}^{n} x_{ij}}$, i=1,…, n, j=1,… n, j=1,…, m

4. Calculate the entropy value of the j-th indicator:

$$e_j = -k\sum_{i=1}^{n} p_{ij}\ln(p_{ij}), \text{ j=1,…, m}$$

Among them, k =$1/\ln(n) > 0$. Satisfy $e_j \geq 0$;

5. Calculate the redundancy degree (difference) of information entropy:

$$d_j = 1 - e_j, \text{ j=1,…, m}$$

Calculate the weights of various indicators:

$$w_j = \frac{d_j}{\sum_{j=1}^{m} d_j}, \text{ j=1,…, m}$$

Calculate the comprehensive scores of each sample:

$$s_j = \sum_{j=1}^{m} w_j x_{ij}, \text{ i=1,…, n}$$

Among them,x_{ij} is the standardized data. The ranking of the weights of the criterion layer [16]: supply chain data = the purchase intention of the target object > the purchasing power of market vegetables = customer feedback > inventory data > competitors' pricing data.

7 Conclusions

This study systematically explores the dynamic sales patterns of vegetable categories and individual items in response to temporal and seasonal variations, proposing a data-driven framework to optimize replenishment and pricing strategies for supermarket profitability. Key findings reveal that leafy and floral vegetables, cauliflower, and aquatic rhizome vegetables exhibit pronounced seasonal trends, while sales of individual items demonstrate long-term fluctuations. The integration of the AMIRA time series model and neural network prediction effectively captures these temporal dependencies, enabling accurate short-term sales forecasting. Furthermore, the identification of inter-category and inter-item correlations through Pearson analysis highlights the interconnected nature of consumer purchasing behavior, offering actionable insights for coordinated inventory management.

The optimization model developed in this study provides a practical tool for supermarkets to balance replenishment quantities and pricing adjustments, ensuring resource allocation aligns with predicted demand. By integrating statistical analysis, machine learning, and operational optimization, this approach transcends conventional heuristic strategies, offering a replicable methodology for perishable goods management. The findings underscore the importance of leveraging temporal data analytics and cross-category correlations in retail decision-making, with implications for reducing waste, maximizing revenue, and enhancing supply chain responsiveness.

This research contributes to the operational management literature by bridging predictive analytics and real-world retail logistics, particularly in perishable goods sectors. Future studies could extend this framework to incorporate external variables such as weather patterns or promotional campaigns, further refining its predictive accuracy and strategic value.

References

[1] Evangelos Theodorou, Evangelos Spiliotis, Vassilios Assimakopoulos, Optimizing inventory control through a data-driven and model-independent framework, EURO Journal on Transportation and Logistics [J], 2023, 12: 100103. https://doi.org/10.1016/j.ejtl.2022.100103

[2] Praveen K. Kopalle, Koen Pauwels, Laxminarayana Yashaswy Akella, Manish Gangwar, Dynamic pricing: Definition, implications for managers, and future research directions [J], Journal of Retailing, 2023, 99(4): 580-593, https://doi.org/10.1016/j.jretai.2023.11.003

[3] Sharon H. Thompson, Rita DiGioacchino DeBate, An Exploratory Study of the Relationship Between Night Eating Syndrome and Depression Among College Students [J], Journal of College Student Psychotherapy, 2009, 24(1): 39-48. https://doi.org/10.1080/87568220903400161

[4] Michael D. Everett, A simplified guide to capital investment risk analysis [J], Planning Review, 1986, 14(4): 32-36. https://doi.org/10.1108/eb054154

[5] Paisit Khanarsa, Arthorn Luangsodsai, Krung Sina piromsaran, Self-ldentification Deep Learning ARIMA [J], Journal of Physics: Conference Series, 2020, 1564: 012004. https://doi.org/10.1088/1742-6596/1564/1/012004

[6] Olu Ola Ogunsote, Bogda Prucnal-Ogunsote, Comfort Limits for the Effective Temperature lndex in the Tropics: A Nigerian Case Study [J], Architectural Science Review, 2002, 45(2): 125-132. https://doi.org/10.1080/00038628.2002.9697500

[7] Steven Shea, AryehD. Stein, Rafael Lantigua, Charles E. Basch, Reliability of the Behavioral Risk Factor Survey in a Triethnic Population [J], American Journal of Epidemiology, 1991, 133(5): 489-500. https://doi.org/10.1093/oxfordjournals.aje.a115916

[8] Aiping Wang, Establishment of Wheat Yield Prediction Mode I in Dry Farming Area Based on Neural Network [J], Neuro Quantology, 2018, 16(6): 768-775. https://doi.org/10.14704/nq.2018.16.6.1654

[9] Siyu Ji, Chenglin Wen, Data Preprocessing Method and Fault Diagnos is Based on Evaluation Function of Information Contribution Degree [J], Journal of Control Science and Engineering, 2018, 2(1): 1-10. https://doi.org/10.1155/2018/6565737

[10] Isabelina Nahmens, Vishal Bindroo, Is Customization Fruitful in Industrialized Homebuilding Industry? [J], Journal of Construction Engineering and Management, 2011, 137(12): 1027-1035. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000396

[11] Zheng Zheng, Bo Liu, Jingwen Gao, Junpeng Li, Xin pei Jiang, Comprehensive evaluation of construction effect of small and medium sized water conservancy projects during construction stage based on FAHP [J], IOP Conference Series: Earth and Environmental Science, 2018, 189(2): 022025. https://doi.org/10.1088/1755-1315/189/2/022025

[12] Guido Fioretti, A mathematical theory of evidence for G. L. S. S hackle [J], Mind & Society, 2001, 2(1): 77-98. https://doi.org/10.1007/BF02512076

[13] Xiaobing Nie, Wei Xing Zheng, Multistability of neural networks with discontinuous non-monotonic piecewise linear activation functions and time-varying delays [J], Neural Networks, 2015, 26(11): 65-79. https://doi.org/10.1016/j.neunet.2015.01.007

[14] Guofa Li, Hongxiang Zhu, Jili Wang, Xinge Zhang, Yongchao Huo, Haiji Yang, Research on importance evaluation of NC machine tool working loads based on AHP-fuzzy comprehensive evaluation method [J], IOP Conference Series: Materials Science and En gineering, 2019, 612(23): 032081. https://doi.org/10.1088/1757-899x/612/3/032081

[15] Shuang Shuang Ma, You Peng Xu, Yu Long Shao, Study on Novel Structures the Coordinated Development between Urbanization and Water Environment in Huzhou City, China [J], Applied Mechanics and Materials, 2012, 209-211: 1040-1047. https://doi.org/10.4028/www.scientific.net/AMM.209-211.1040

[16] Zhijie Jia, Jianbing Peng, Quanzhong Lu, Penghui Ma, A Comprehensive Method for the Risk Assessment of Ground Fissures: Case Studyofthe Eastern Weihe Basin [J], Journal of Earth Science, 2023, 34(6): 1892-1907. https://doi.org/10.1007/s12583-022-1799-6