Research on Pricing and Replenishment Decision of Vegetable Products Based on Optimization Algorithms

Huan Liu*

School of Electrical and Information Engineering, Anhui University of Technology, Ma'anshan, China, 243002

*Corresponding author: 15695557069@163.com

Abstract. With the improvement of people's living standards, people's requirements for the freshness and variety richness of food are also constantly increasing. Due to the short shelf life of fresh vegetable products, reasonable prediction and formulation of pricing and replenishment strategies have a significant impact on supermarket revenue. This article first cleans the data and uses Excel and Grabbs statistics to determine the missing and abnormal values of the data. The missing values are supplemented by 0, and the abnormal values are determined and eliminated through manual intervention. Conduct descriptive statistics on the preprocessed data to obtain partial data features of categories classified by day and month. Analyze numerical features to determine the distribution pattern of data using monthly classification, and then use SPSS PRO to conduct Spearman correlation analysis to obtain the relationship between vegetable categories. Next, this article quantifies six data indicators, including daily sales volume of each category, average unit price during normal sales, and discount degree. Then, MATLAB is used for regression fitting analysis to obtain the relationship expressions between each indicator and the total sales volume. Excel analysis is used to summarize the data to obtain the available variety information of each item of vegetables in the supermarket from June 24th to 30th, And use Python software to establish a time series prediction model to analyze and process the information of supermarket products, and obtain the demand for individual products on July 1st. Using the single product selected by the supermarket as the 0-1 decision variable, and the total operating profit and supply demand relationship of the supermarket as the dual objective variables, a multiple linear regression model is constructed to obtain the final replenishment policy and pricing decision. The data in this paper is from the National College Students Mathematical Modeling Competition C question.

Keywords: Spearman Correlation Analysis, Time Series Prediction Model, Multiple Linear Regression Model.

1. Introduction

In the sales of fresh food supermarkets, taking vegetable products as an example, if most varieties are not sold on the same day, they cannot be sold the next day. Therefore, supermarkets need to make daily replenishment decisions based on the historical sales and demand of products. Due to the large variety and different origins of vegetables, and the fact that supermarkets usually purchase vegetables between 3:00 and 4:00 in the morning, merchants are not clear about the specific individual products and purchase prices when making replenishment decisions. Therefore, the supermarket adopts a "cost plus pricing" method to price, offering discounts on goods with transportation losses and poor phase change. From the perspectives of demand and supply, it is found that vegetable products are related to time, and the variety of supply is relatively abundant from April to October. The sales space of supermarkets is limited, so a reasonable sales mix is very important. In order to make accurate replenishment and pricing decisions, reliable market analysis is required.

2. Data cleaning based on Grabbs criterion

2.1. Data preprocessing

For outliers, this article uses the Grubbs criterion [1] to analyze the data. The Grabbs criterion is a commonly used statistical method for detecting outliers in a dataset. It is based on the principles of

statistics and determines whether there are outliers by calculating the difference between outliers and the average value. In data analysis, this article divides the Grabbs criterion into two steps: first, calculate the difference (numerator) between each value of the data and the sample mean, and then calculate the standard deviation (denominator) of all data values. Compare the difference between outliers and the mean with the standard deviation to obtain the Grabbs statistic. Based on the size of the Grabbs statistic, a critical value (calculated at the significance level) can be applied to determine whether there are outliers. If the Grabbs statistic of outliers is greater than the critical value, it can be determined as an outlier. It was found that the sales date was June 9, 2022, the scan code sales time was 09:31:57.045, and the sales volume of the "Zongye (Bag) (1)" product with a single item code of 102900011034354 was 160kg. Manual intervention identified it as abnormal data and deleted it. For missing values, this article uses the function library in Excel to search for missing values. If there are missing values, interpolation method is used to fill in the missing values. Upon checking the attachment data, it was found that there are a large number of blank values in the data of this question. As the product has not been purchased and the value is reasonable, it is sufficient to fill in 0. This article uses two methods of data aggregation: daily aggregation and monthly aggregation to analyze the data.

2.2. Descriptive Statistical Index of Data

This article seeks to find the distribution patterns and interrelationships of sales volume of various categories and individual products of vegetables by calculating descriptive features such as mean, range, variance, kurtosis, coefficient of variation, and quartiles of aggregated data:

(1) Mean

The mean can describe the central trend of the dataset and represent the overall level of the dataset $\overline{x}$

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

For daily aggregation, represents the average sales volume of the product day; For monthly aggregation, represents the average sales volume of the product month. $\overline{x}$, i, n, $\overline{x}$, i, n.

(2) Kurtosis

Kurtosis is a statistic that measures the sharpness or flatness of a set of data distributions. It describes the peak or low peak characteristics of data distribution, which can understand the sharpness of the data near the central position: K

$$K = \frac{1}{n}\sum_{i=1}^{n}\left[\left(\frac{x_i - \overline{x}}{s}\right)^4\right] \tag{2}$$

(3) Coefficient of variation

The coefficient of variation can measure the relative degree of variation of a dataset. The smaller the numerical value, the smaller the relative variability of the dataset, while the larger the numerical value, the greater the relative variability of the dataset: CV

$$CV = \frac{s}{\overline{x}} \times 100\% \tag{3}$$

(4) Quartile

The quartile is an indicator in statistics that describes the distribution of a dataset. It divides a dataset into three equal parts, namely the first (less than 25% of data), the second (median), and the third quartile (less than 50% of data). By using quartiles, one can understand the distribution of the dataset at different positions and the relative position of the mode (second quartile) in the dataset distribution.

(5) Cost plus pricing[2]

The basic idea of the cost plus pricing method is to set product prices based on the unit cost of the product plus a certain proportion of profits. Basic formula price=unit cost (l+cost profit margin)

3. Correlation analysis and descriptive statistics based on Spearman coefficient

3.1. Daily aggregation of vegetable categories and individual products

The large data prediction model for the user's electricity consumption is implemented in the Clementine software.

3.2. Analysis of experimental results

Table 1: Descriptive Characteristics of Selected Vegetable Categories Clustered by Day

Category	Mean	Kurtosis	Coefficent	Quartile (25%, 50%, 75%)		
Cauliflower	38.53	0.689	0.588	14.291	21.892	30.374
Flowers and leaves	182.969	0.475	0.471	30.408	45.853	68.498
Peppers	84.413	0.771	0.633	10.067	14.674	29.953
Solanula	21.364	0.755	0.606	5.95	18.262	36.382
Food fungus	70.126	0.877	0.691	9.437	18.253	27.202
Aquatic rhinomes	37.402	1.146	0.838	3.606	7.385	18.806

According to Table 1, it can be seen that daily clustering, mean, kurtosis, range, coefficient of variation, etc. have significant fluctuations and are not suitable for analyzing the relationships between different categories. Therefore, this article uses monthly statistical data[3] for analysis.

3.3. Monthly aggregation of various vegetable categories

This article summarizes the data on a monthly basis, using months as a unit. It is calculated based on 36 months and is summarized into 36 pieces of data. And use MATLAB software to further analyze the collected data to obtain descriptive characteristics such as mean, range, variance, kurtosis, coefficient of variation, quartiles, etc. of the total sales volume of each vegetable category. The specific data is detailed in Table 2.

Table 2: Descriptive Characteristics of Partial Categories of Vegetables Clustered by Month

Category	Mean	Kurtosis	Coefficent	Quartile (25%, 50%, 75%)		
Cauliflower	1160.179	0.511	0.384	23.305	34.072	48.61
Flowers and leaves	5514.472	0.464	0.312	128.644	173.192	223.553
Peppers	2544.129	0.257	0.415	51.219	72.925	102.172
Solanula	623.105	0.164	0.456	12.226	18.87	27.142
Food fungus	2113.52	0.265	0.459	39.245	57.535	89.254
Aquatic rhizames	1127.26	0.97	0.573	15.502	30.194	50.906

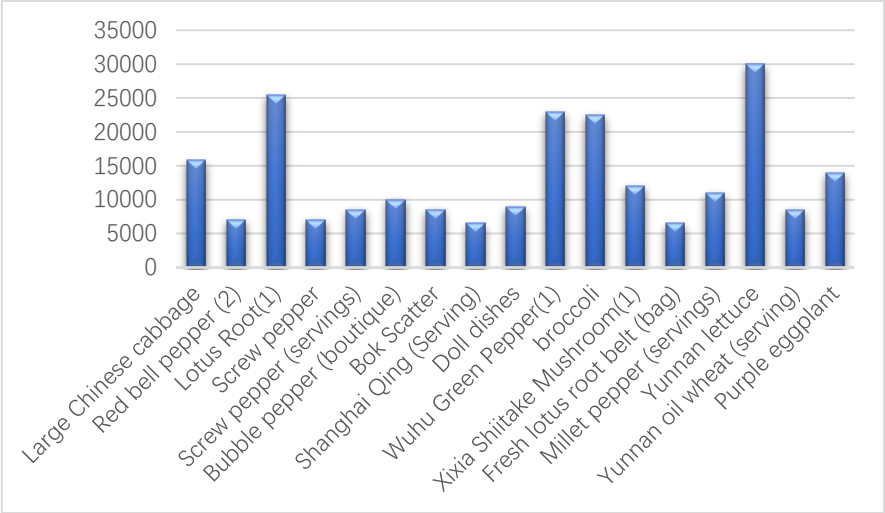

Fig 1: Monthly Clustering Column Chart of Sales Volume of Some Single Products

According to bar chart 1, the top ten single item sales fig 1 show that Yunnan lettuce has had the highest sales volume in the past two years, followed by Wu hu green pepper. The popularity level of the two single items is relatively high.

In order to observe the changes in sales volume between different categories of vegetables more intuitively, a line chart[4] was drawn using Python, as shown in Fig 2:

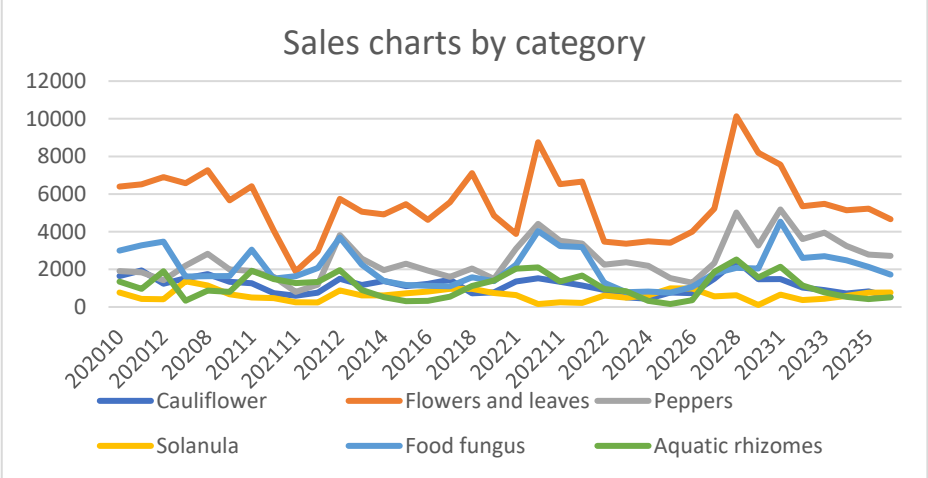

Fig 2: Line Chart of Sales Volume Distribution by Monthly Category

Each variety of vegetables has a maximum and minimum value throughout the year, with significant fluctuations in floral and leafy varieties, with a maximum value appearing every month or so, possibly due to external factors such as promotions. However, the fluctuation range of eggplants is not significant, indicating that seasonal climate or months have little impact on their growth.

Import the monthly aggregated data into SPSSPRO software for analysis, and obtain the correlation coefficients between various categories of vegetables through Spearman correlation analysis ρ.

$$\rho = \frac{\sum_i \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_i \left(x_i - \bar{x}\right)^2 \sum_i \left(y_i - \bar{y}\right)^2}} \tag{4}$$

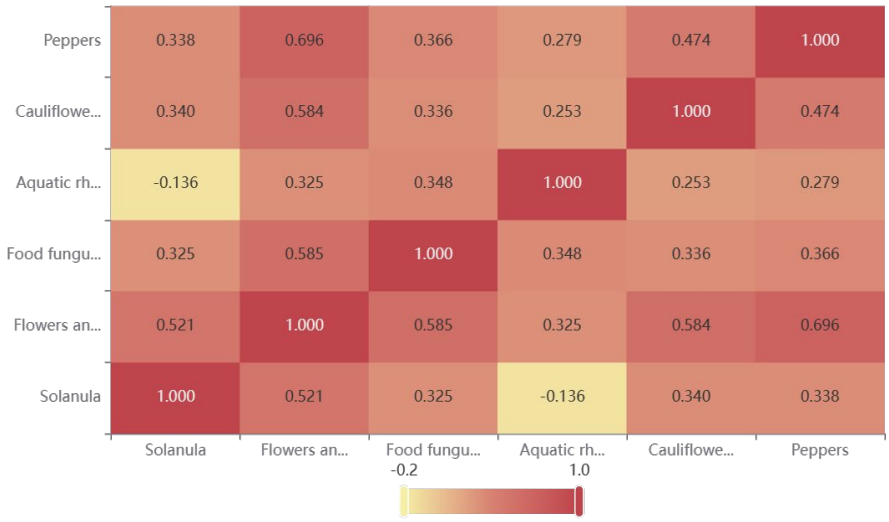

Fig 3: Heat map of correlation numbers between different categories of vegetables

Based on the above data analysis results and Figure 3, it is easy to obtain the distribution pattern of various categories of vegetable products and the correlation between each category. Among the categories, there is a strong correlation between flowers, leaves, and cauliflower, which may be purchased at the same time during sales. There is a negative correlation between edible mushrooms and eggplants, aquatic rhizomes, and eggplants. Consumers do not tend to purchase this combination at the same time when purchasing, and it is possible that the two have different market times, Different growth cycles[5].

4. Prediction model based on regression fitting equation

4.1. Quantification of data

In order to facilitate the analysis of the relationship between total sales volume and cost plus pricing, and to predict replenishment and pricing strategies for the next week, this article quantifies the data into six indicators on a daily basis:

(1) Total daily sales of each category

$$x_a = \sum_{b=1}^{m} x_{ab} - \sum_{b=1}^{m} x'_{ab} \tag{5}$$

x_a Represents the total sales volume of a certain type of vegetables on the day, represents the number of types of vegetables of the same type, represents the total sales volume of all vegetables of the same type on the day, and represents the return volume of all vegetables of the same type on the day a, b, x_{ab}, a, x'_{ab}, a

(2) Average unit price during normal sales

$$\omega_t = \frac{\sum_{b=1}^{m} y_{ab}\omega_{ab}}{\sum_{b=1}^{m} y_{ab}} \tag{6}$$

y_{ab} Represents the normal sales volume of the first vegetable on day one, and represents the corresponding sales unit price. a, b, ω_{ab}.

(3) Discount average unit price

$$v_a = \frac{\sum_{b=1}^{m} y'_{ab} v_{ab}}{\sum_{b=1}^{m} y'_{ab}} \tag{7}$$

y'_{ab} Represents the marketing volume of the first vegetable on the day after the discount, and represents the corresponding discounted unit price. This indicator reflects which categories of vegetables are not sold out and has an impact on replenishment decisions.a, b,.v_{ab}

(4) Discount rate

$$\lambda_a = \frac{\omega_a}{v_a}\% \tag{8}$$

(5) Average wholesale price

$$p_a = \frac{\sum_{b=1}^{m} (x_{ab} - x'_{ab}) p_{ab}}{x_a} \tag{9}$$

p_{ab} Represents the wholesale price of the first vegetable on day one. The cost can be determined by wholesale prices.a, b

(6) Transport loss rate

$$s_a = \frac{\sum_{b=1}^{m} (x_{ab} - x'_{ab}) s_{ab}}{x_a} \tag{10}$$

4.2. Establishing Fitting Equations

Perform correlation analysis on the above six indicators to obtain the relationship between each indicator and the total sales volume[5]. Use MATLAB to fit to obtain Tables 3, 4, 5, 6:

Table 3 : Total sales volume and time of various types

Category	Fit the equation	R
Cauliflower	$x = -5.1927a^2 + 110.88a - 569.05$	0.3889
Flowers and leaves	$x = -42.564a^3 + 673.48a^2 - 4712.7a + 12348$	0.039
Chili pepper	$x = -1.3778a^4 + 68.065a^3 - 1254a^2 + 10215a - 31014$	0.2511
Solanula	$x = -1.3778a^4 + 68.065a^3 - 1254a^2 + 10215a - 31014$	0.1499
Food fungus	$x = -2.0125a^4 + 112.62a^3 - 2351.5a^2 + 21713a - 74772$	0.2079
Aquatic rhinomes	$x = -0.0009a^4 - 0.067a^3 + 1.8788a^2 - 21.329a - 126.26$	0.1072

Table 4: Fitting relationship between partial total sales and discounted unit price[6]

Category	Fit the equation
Chili pepper	$x = 20.623v^2 - 471.62v + 1940.2$
Aquatic rhizames	$x = 0.0219v^2 - 0.6749v + 6.3131$
Food fungus	$x = 0.0691v^2 - 1.841v + 16.931$
…	…

Table 5: Fitting relationship between total sales volume and average unit price of various types

Category	Fit the equation
Solanula	$x=0.3959\omega^2-8.6192\omega+63.591$
Flowers and leaves	$x=3.2319\omega^2-49.985\omega+343.59$
Chili pepper	$x=0.4239\omega^2-11.467\omega+10.007$
Hana	$x=0.5229\omega^2-14.256\omega+156.92$
Aquatic rhinomes	$x=-0.0137\omega^2+2.402\omega+10.161$
Food fungus	$x=-0.5618\omega^2+6.8993\omega+49.951$

Table 6: Fitting relationship between total sales of various types and average wholesale unit price

Category	Fit the equation
Aquatic rhizames	$x=0.4591p^2-15.661p+139.42$
Food fungus	$x=0.9038p^2-21.866p+160.36$
Solanula	$x=0.2316p^2-3.969p+36.936$
Chili pepper	$x=4.6658p^2-71.586p+332.1$
Flowers and leaves	$x=5.4023p^2-78.028p+411.2$
Hana	$x=-0.0021p^2+1.6574p+1.3648$

5. multi-objective linear programming problem based on time prediction model

5.1. Model preparation

Aggregate and summarize the data through Excel analysis pivot table to obtain the available variety information of various vegetable items in the supermarket from June 24th to 30th. At the same time, using Python software to use a time series prediction model [7][8] to analyze and process the information of supermarket products, the demand for each item of supermarket vegetables on July 1st was obtained, as shown in Table 7.

Table 7: Forecast of Demand for Single Product Sold on July 1st

Item coding	Demand for each single product of vegetables on July 1
102900005115250	4.624142857142857
102900005115762	8.923
102900005115779	3.418045545514443
102900005115786	13.296714285714286
102900005115823	3.8612857142857138
.....	.....
.....	.....
106971533450003	8.857142857142856
106949711300259	14.207373017705232
102900051000944	3.930977825329259
102900051000463	1.2405000000000004

5.2. Model Establishment

(1) Determine decision variables

The decision variable is the single product selected by the supermarket, and a 0-1integer variable is introduced: h_i , If you order the I th item, h_i takes 1, if you do not order the product, h_i takes 0

(2) Establish optimization goals

Firstly, by analyzing the quantitative relationship between sales unit price and discounted sales unit price, it can be concluded that

$$\begin{cases} m_i = f(v_i) \\ n_i = g(v_i) \end{cases} \tag{11}$$

The optimization target is the total operating profit[6] of the supermarket Z:

$$Z = \sum K_i v_i m_i (1-\theta_i) + \sum K_I v_i n_i \theta_i - \sum K_i v_i c_i \tag{12}$$

The total number of vegetable items is N, and the corresponding sales prices for each item are recorded. The discounted sales unit price is, the demand for vegetable items is, the transportation loss rate of vegetable items is, and the wholesale price[9] of vegetable items represents the order quantity of each item $d_1, d_2 ... d_N$ $m_i (i=1,2,...,N)$, $n_i (i=1,...,N)$, $K_i (i=1,...,N)$, $\theta_i (i=1,...,N)$,. $c_i (i=1,...,N)$,. v_i .

(3) Establish constraints

a. The supply and demand relationship of each individual vegetable product is:

$$g_i = \sum |v_i - \partial_i| \tag{13}$$

b. Control the total number of available vegetable items between 27 and 33

$$\begin{cases} 27 \leqslant \sum h_i \leqslant 33 \\ v_i \geqslant 2.5 \end{cases} \tag{14}$$

c. The order quantity of each item shall not be less than the display quantity of 2.5kg

$$x_i \geqslant 2.5 (i=1...N) \tag{15}$$

In summary, the following multivariate linear programming models can be obtained

$$\max Z = \sum k_i v_i m_i (1-\theta_i) + \sum K_i v_i n_i \theta_i - \sum K_i v_i c_i \tag{16}$$

$$\min g_i = \sum |v_i - \partial_i| \tag{17}$$

$$s.t. \begin{cases} 27 \leqslant \sum h_i \leqslant 33 \\ x_i \geqslant 2.5 (i=1...N) \end{cases} \tag{18}$$

5.3. Solution of the model

By using MATLAB software to solve the replenishment policyand pricing decision of the above multiple linear programming model[10], as shown in Table 8

Table 8: Supermarket Replenishment Policy and Pricing Decision Table

Item coding	Total Daily Replacement (KG)	Pricing Strategy%	Item coding	Daily Replacement Total Quantity (KG)	Pricing Strategy (%)
102900005115625	3.172	111.79%	102900011035078	2.762	102.07%
102900005115779	10.443	134.23%	102900011035740	1	85.37%
102900005115878	9.014	142.23%	102900011021842	7	71.14%
102900005115861	18.215	94.02%	102900011023464	5.317	127.20%
102900005116257	2.299	46.21%	102900011030059	13	98.47%

6. Conclusion

This article mainly explores the distribution patterns and interrelationships of various categories and individual items of vegetables, as well as the replenishment and pricing strategies of supermarkets during the corresponding time periods, and predicts future strategies. Pre processing of data reduces the impact of outliers and missing values on the results, and improves the accuracy of the model. In the process of optimizing the supermarket replenishment plan, this article mines the data at each step and uses a pivot table to quickly and intuitively simplify a large amount of data. In the exploration, dynamic programming models and multi-objective linear programming models are used to scientifically and reasonably determine the daily replenishment total amount and pricing strategy of the supermarket, with high reliability.

References

[1] Li Huaixiang, Zhuo Zhihai. Improved 2D CA-CFAR based on Grubbs criterion [J]. Journal of Beijing University of Information Technology: Natural Science Edition, 2023, 38

[2] Wang Zirun, Wang Qianyi, Wang Xiaomeng, Yuan Wenjia, Wei Xiao, Liu Wei, Fan Xiuzhu. Research on supplier selection strategy based on dynamic weighting method from the perspective of returning home employment [J]. Technology and Market, 2023,30 (07)

[3] Cao Yili, Gao Min, Sun He, et al. A dynamic reactive power compensation device loss calculation method based on regression prediction: CN202211362059.6 [P]

[4] Xu Wenting, Xu Liqun: A Study on the Optimal Combination Strategy of Profits for Volatile Products Influenced by Multiple Factors [J]. Shanghai Management Science 2016

[5] Li Yuan. Research on Dual Channel Retail Pricing and Inventory Joint Optimization Considering Reference Price Effect [D]. Yanshan University, 2022

[6] Zhao Ling, Liu Zhixue. Joint replenishment and pricing strategy considering customer returns and fixed costs [J]. Operations Research and Management, 2022

[7] Cibicong, Zhang Pinyi. Financial Time Series Prediction Based on ARIMA-LSTM Model [J]. Statistics and Decision, 2022 (11): 5

[8] Yu Shixin, Yao Zejia, Wang Haodong, et al. Bank loan strategy based on BP neural network and linear programming model [J] two thousand and twenty-one

[9] Wu Lifang. A multi-stage pricing model for fresh agricultural products considering freshness and consumer utility [J]. Comprehensive Transportation, 2022,

[10] Li Jiaxu. A Study on the Pricing of Dual Channel Air Tickets under Demand Orientation [J]. Logistics Technology, 2021, 40 (7): 8