Research on Automatic Pricing and Replenishment Decisions for Vegetable Commodities Based on Pearson Correlation Analysis and VAR Model

Jiahe Chen [1, #, *], Jingwei He [2, #] and Qiyu Mao [2, #]

[1] School of Electronic Information and Automation, Tianjin University of Science and Technology, Tianjin, China

[2] School of Mechanical Engineering, Tianjin University of Science and Technology, Tianjin, China

* Corresponding author: 3403067454@qq.com

[#]These authors contributed equally to this work

Abstract. This paper investigates the automatic pricing and replenishment decisions for vegetable commodities based on Pearson correlation analysis and the Vector Autoregression (VAR) model. By analyzing vegetable sales data, box plots and histograms reveal the sales volume distribution patterns of different categories and individual vegetable products, and Pearson correlation analysis uncovers the intrinsic relationships among them. A VAR model is further constructed to forecast the sales volume and pricing of vegetables, combined with the cost-plus pricing strategy from economic principles to determine the replenishment volume and pricing strategy that maximizes profits. The study demonstrates that the model can effectively guide supermarkets in achieving optimal pricing and replenishment for vegetable products, enhancing the economic benefits.

Keywords: MATLAB, Pearson Correlation Analysis, STC function, TR function, VAR model.

1. Introduction

In the fast-paced retail market of today, vegetables, as daily consumer goods, have a significant impact on the operational efficiency and profits of supermarkets [1]. With the development of big data and machine learning technologies, traditional experience-based pricing and replenishment methods are gradually being replaced by more scientific and precise models. This paper aims to explore an automatic pricing and replenishment decision-making method for vegetables based on Pearson correlation analysis and the Vector Autoregression (VAR) model, with the expectation of providing supermarket managers with a new decision support tool [2, 3].

Firstly, this paper analyzes vegetable sales data, using box plots and histograms to reveal the distribution patterns of sales volume for different categories and individual vegetable products. Pearson correlation analysis further uncovers the intrinsic relationships among them [4]. These analyses lay the foundation for constructing a more accurate sales forecast model. Subsequently, a VAR model is constructed in this paper to predict the sales volume and pricing of vegetables, combined with the cost-plus pricing strategy from economic principles to determine the replenishment volume and pricing strategy that maximizes profits [5, 6]. The study demonstrates that the model can effectively guide supermarkets in achieving optimal pricing and replenishment for vegetable products, enhancing economic benefits.

2. Data processing

2.1. Data analysis and processing

First of all, the data are cleaned, the four annexes are combined with the single product code as a fixed value, and the abnormal values, missing values, and duplicate values are checked and processed, and at the same time, for the sales flow detail data, it is classified according to the category and single product, and the sales volume of each category and single product is calculated. To ensure the accuracy of the model results and make them more relevant.

2.2. Distribution pattern of sales volume of each category and single product of vegetables

For the distribution law of the sales volume of each category of vegetables, it is visually represented by box plots (box plots are a kind of statistical charts that use five statistics in the data: the minimum value, the upper quartile, the median, the lower quartile, and the maximum value to describe the data). This is shown in Fig. 1.

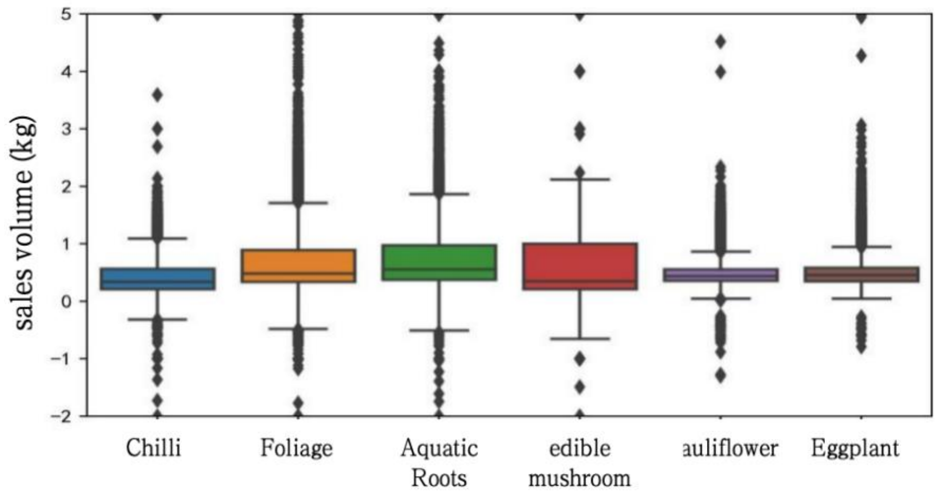

Figure 1. Vegetable sales by category box plot

We can see very intuitively that the box plot edibles and foliage categories have fewer discrete points, the mean value is large, indicating that it has a large amount of sales data, high sales data value, cauliflower and eggplant categories on the contrary less, poorer sales, and may be less popular with customers. For individual categories, we use excel's filter sorting to sort their sales (see Annex 1), and take the top ten sales to make a characterization as shown in Table 1.

Table 1. Vegetable Unit Sales

Single category	Summation term: sales volume (kg)
Wuhu Green Pepper (1)	28164.33
Broccoli	27537.23
Net Lotus Root (1)	27149.44
Chinese Cabbage	19187.22
Yunnan Lettuce	15910.46
Golden Needle Mushroom (box)	15596.00
Yunnan Lettuce(portion)	14325.00
Purple Eggplant (2)	13602.00
Xixia Shiitake Mushroom (1)	11920.23
Millet Pepper (portion)	10833.00

3. Correlation Analysis

3.1. Pearson Correlation Analysis

For the correlation analysis of each category of vegetables as well as between individual products, we used the Pearson correlation analysis model to analyse the degree of correlation of their sales volume.

The Pearson linear correlation coefficient is the most commonly used linear correlation coefficient. The most applicable form of data: linear data, continuous and in line with normal distribution, the difference between the data cannot be too large. There are m objects and n columns of indicators, which can form the data matrix X $(xij)_{m\times n}$. Now study the correlation between column a xa and column b, in the data matrix. column bathe correlation, let the correlation coefficient of column a, b is rho (a, b), then there are:

$$rh(a,b)=\frac{\sum_{i=1}^{m}(x_{a,i}-\overline{x_a})(x_{b,i}-\overline{x_b})}{\sum_{i=1}^{m}(x_{a,i}-\overline{x_a})^2\sum_{j=1}^{m}(x_{b,j}-\overline{x_b})^2}. \tag{1}$$

Where m is the length of each column and the correlation coefficient ranges from -1 to 1. A value of -1 indicates a perfect negative correlation, while a value of +1 indicates a perfect positive correlation. A value of 0 means that there is no correlation between the indicators.

3.2. Correlation Between Categories and Individual Products

Based on the processed data, the correlation coefficient p was calculated by python, and the heat map of correlation between categories and individual products of vegetables was made.

Table 2. Three-line table of correlation coefficients for each vegetable category

cd	Aquatic Roots	Foliage	Cauliflower	Eggplant	Chilli	edible mushroom
Aquatic Roots	1.000	0.561	0.542	0.106	0.614	0.670
Foliage	0.561	1.000	0.626	0.323	0.659	0.631
Cauliflower	0.542	0.626	1.000	0.348	0.551	0.522
Eggplant	0.106	0.323	0.348	1.000	0.318	0.141
Chilli	0.614	0.659	0.551	0.318	1.000	0.687
edible mushroom	0.670	0.631	0.522	0.141	0.687	1.000

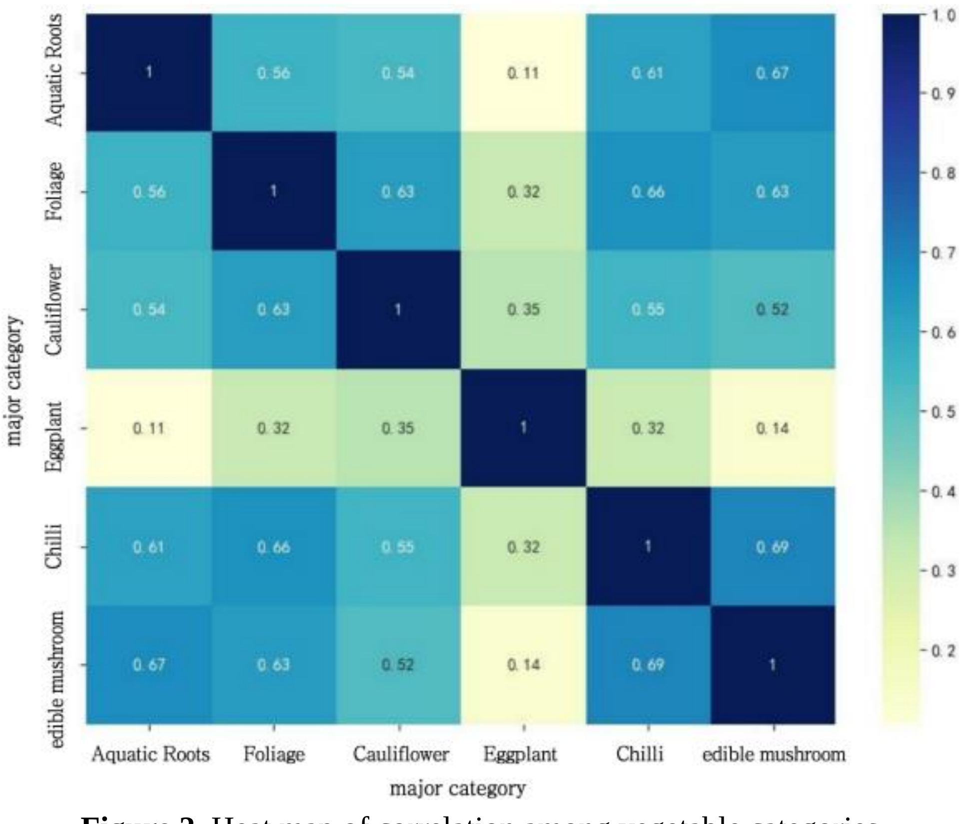

Figure 2. Heat map of correlation among vegetable categories

As can be seen from Table 2 and Fig. 2, there is significant variability in correlation among vegetable categories. Among them, there is a strong correlation between leafy and flowering vegetables, hot pepper and edible mushroom.

4. VAR Model

4.1. Model Establishment

We know the loss rate of each order, and we can use its relationship with sales volume to calculate the cost of various types of vegetables.

$$Total = Q_{sell} \times (1 + Damage)\,. \tag{2}$$

$$Q_{sell} = Q_1 + Q_2 + Q_3 + Q_4 + Q_5 + Q_6\,. \tag{3}$$

$$Cost = (Q_x \times 1 + Loss) \times P_{wholesale}\,. \tag{4}$$

For the relationship between the total sales volume of vegetables and the cost-plus pricing, we first used MATLAB tools to express the functional relationship between the total sales volume of vegetables and the cost-plus pricing and profit, and used Polyfit tools to fit each variable. Under the situation of an ideal perfect competition market, STC model is roughly cubic curve, TR model model is roughly a model, so the initial basic setup is as follows:

$$STC = aQ^3 + bQ^2 + cQ + d\,. \tag{5}$$

Where a, b, c, d are the coefficient constants of the STC function model, and f are the function constants of the TR model, and the processed data needs to be brought into the forward line for fitting.

In the short term, as the limit of overselling space has always existed, the scale is regarded as stable. Profit maximization can be achieved by adjusting the total amount and price of each vegetable category. The basic idea is shown in Fig.3.

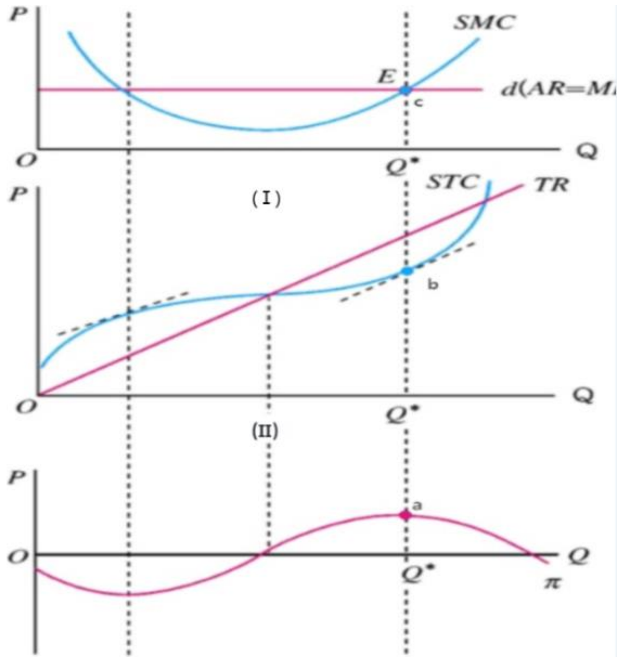

Figure 3. Profit maximization equilibrium condition

On this basis, we use python to establish a vector autoregressive model (VAR) to predict the total sales volume and pricing of each vegetable category in the following 7 days, and compare them with the price and total volume under the condition of profit maximization, so as to determine the daily replenishable volume and pricing strategy of each vegetable category in the coming week.

4.2. Automatic Pricing and Replenishment Decisions

With cost as the dependent variable (y) and sales volume as the independent variable (x), matlab plotify was used to fit the STC function of sales volume and cost of six vegetable categories, most of the original data fall on the fitted curve, because the profit maximization in the economic principle we use mainly focuses on the STC slope, that is, the situation of SMC=AR=MR=P, so the STC fits mainly the trend relationship between quantity and cost.

Table 3. Initial STC fitting model

Class	RMSE
Aquatic rhizomes	132.79
Anthophyllum	15.6048
Cauliflower	0.725
Solanacea	0.3878
Capsicum	0.4426
Edible mushroom	6.2757

In Table 3, we can see that for the fitting model of Mosaic and edible fungi, RMSE were 15.6048 and 6.2757 respectively, which had a large deviation compared with the fitting of actual data. For the functional fitting model of aquatic rhiza, the RMSE of the model fitting evaluation result was 132.79, and the root-mean-square value of the difference between the predicted value and the true value was too large, so the fitting was continued.

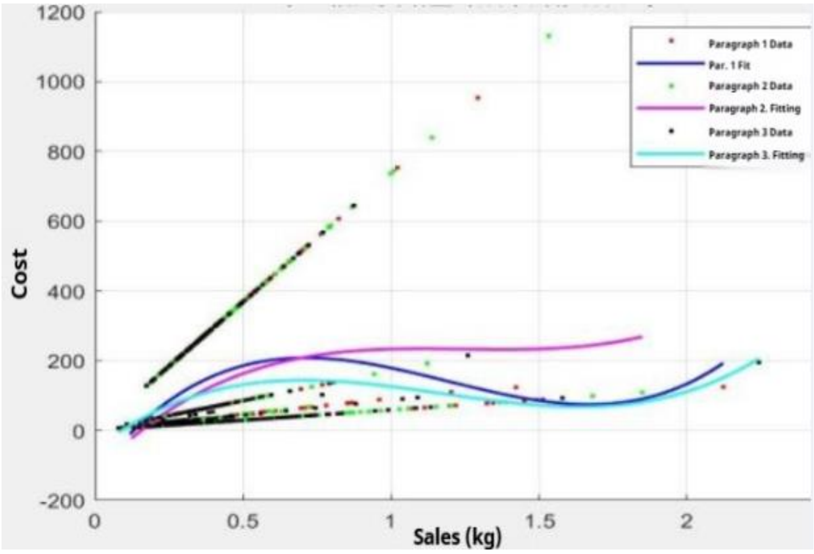

Figure 4. Piecewise fitting function of sales volume and cost of aquatic roots

By dividing the data of aquatic rhizome into two paragraphs, two fitting lines can be found as shown in Fig. 4. We observe the fitting of paragraph 1 and paragraph 2, and find that paragraph 2 can better reflect the trend of total quantity and cost. We take the function model of paragraph 2 as the fitting function of aquatic rhizome. That is, y = 187.8266 x^ 3-659.8134 x^2 + 650.3163 x - 52.4538, and then this method is used to modify other models.

Next, with income as the dependent variable (y) and sales volume as the independent variable (x), matlab was used to calculate the income and sales volume of 6 types of vegetables. Row fitting, using matlab to extract the subset data of the major class, and then extract the sales and income data, using polyfit function to get a multinomial. The TR model fitted by formula. In the fitting of edible fungi, it can be seen from its TR fitting diagram that there are obviously two ends of the data, and we choose the data with a positive slope, which is not only in line with the ideal model of economics, but also easy to calculate.

It can be seen from the idea in Fig. 4 that profit maximization can be achieved when the slope of STC image is equal to the slope of TR for the second time. This idea is used to differentiate STC, and the result is compared with the slope of TR. Finally, the optimal weight of six vegetable categories, namely the average price PR, can be obtained when profit maximization is achieved. The results are shown in Tables 4 and 5.

Table 4. Optimal weight of vegetables obtained by STC and TR

Class	Weight (kg)
Aquatic rhizomes	[0.683612706586088, 1.65831117426998]
Anthophyllum	[0.668632841811591]
Cauliflower	[0.142648899312970, 1.90142253526538]
Solanacea	[3.35988938053097]
Capsicum	[-0.229666875050064, 0.797765506995436]
Edible mushroom	[-0.155055179825895,0.610823117004523]

Table 5. Three Scheme comparing

Class	Sales unit price (Yuan/kg)
Aquatic rhizomes	13.92072667
Anthophyllum	5.462580736
Cauliflower	7.072015522
Solanacea	8.272471191
Capsicum	10.0815812
Edible mushroom	15.23062914

4.3. Model Evaluation Model evaluation

Using the python editor, after importing relevant data, draw the sales volume and sales price time series of each vegetable category, and observe their time series changes. As shown in Fig. 5, the sales volume and selling price of the 6 vegetable categories have a series of fluctuations in the time series as a whole.

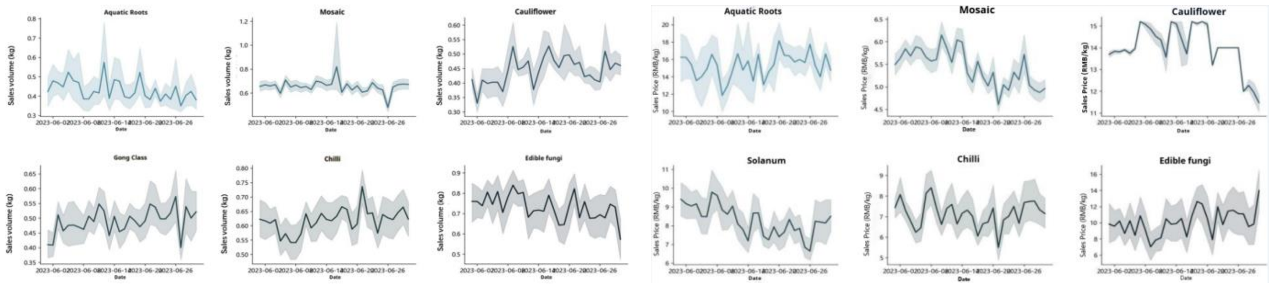

Figure 5. Time series of sales volume and prices of 6 vegetable categories (2023.6)

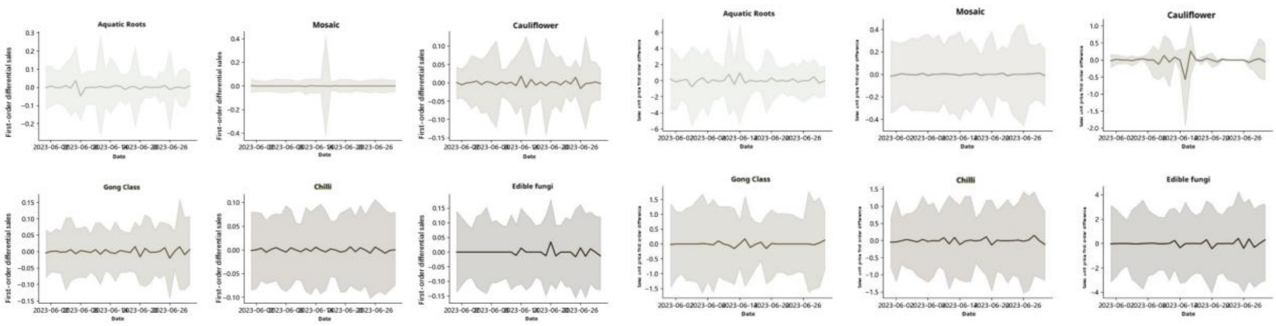

Figure 6. First-order differential timing of sales and unit price of 6 vegetable categories

After the first-order difference of the relevant data, the time series diagram is drawn, as shown in Fig. 6. It is assumed that the first-order difference series is a stationary series for sales volume and sales price. In order to verify the conjecture, the stationarity test and white noise test are conducted for sales volume and sales price as shown in Table 6 and Table 7.

Table 6. Stationarity test

Data index	ADF value	P value (significant level 0.05)
Sales volume	-30.65720647982173	0.0*
Unit sales price	-38.914982366236465	0.0*

Table 7. White noise test

Data index	Ljung-Box statistic	lb_pvalue (P-value)
Sales volume	4264.908167	0.0*
Unit sales price	4264.908167	0.0*

5. Conclusion

This study has successfully developed an automatic pricing and replenishment decision-making model for vegetable commodities, leveraging Pearson correlation analysis and the Vector Autoregression (VAR) model. The model has been applied to analyze sales data, uncovering the sales volume distribution patterns and intrinsic relationships among different categories and individual products. The findings from the correlation analysis have been instrumental in establishing a VAR model that forecasts sales volume and pricing, which is then integrated with a cost-plus pricing strategy to determine optimal replenishment volumes and pricing strategies. The results indicate that the proposed model can significantly enhance supermarket operations by providing data-driven insights for pricing and replenishment decisions. By maximizing profits through optimized pricing and inventory management, supermarkets can improve their economic benefits and operational efficiency.

References

[1] SHE ZHANG, XU JIAHUA, YAO ZHIYU, et al. Analysis of online shopping big data based on Pearson's correlation coefficient--Taking the transaction records of Bai Runju flagship shop on Tmall as an example [J]. Journal of Hanshan Normal College, 2020, 41 (03): 16 - 22.

[2] SUN Zhizhi, WANG Fayuan. Correlation analysis of price volatility of beef, chicken and pork - VAR model analysis based on national data from 2006 - 2021 [J]. Agricultural Outlook, 2023, 19 (06): 24 - 32.

[3] Li, Hua-na. Analysis of domestic crop price forecasting by VAR model [D]. Chongqing University, 2022. DOI: 10.27670/d. cnki. gcqdu. 2022. 004086.

[4] A Qi. Research on the Pricing of Scientific and Technical Research Services Based on Cost-plus Pricing Method [J]. Library Research and Work, 2021, (10): 25 - 31+24.

[5] HAN Junhua, Gan Shengdao. A review of cost-plus pricing method [J]. Finance and Accounting Monthly, 2012 (22): 74 - 75. DOI: 10.19641/j.cnki.42 - 1290/f.2012. 22. 034.

[6] Li Qilin. Post-parallelization era: How policy affects market interest rates [J]. New Wealth Management, 2019 (10): 41 - 47.