Study on the Relationship Between Vegetable Pricing and Sales Volume Based on the Data-Driven Approach

Zhicai Ao, Shuangxu Li *, Yufeng Du, Guodong Wang, Rui Liu

School of Mechanical Engineering, Inner Mongolia University of Science & Technology, Baotou, China, 014000

* Corresponding Author Email: shuangxuli@126.com

Abstract. The study of the relationship between pricing and sales volume of vegetables is important for optimizing supply chain management and enhancing the profitability of agricultural products, and has become a key issue in responding to market competition and the diversification of consumer demand. The relationship between different categories of vegetables, and between them and sales volume is currently unclear. For this purpose, this study analyzed the vegetable categories using descriptive statistical analysis and linear regression analysis and aggregated them with the help of Spearman's correlation test. This study will explore the relationship between sales volume and cost pricing to make the final pricing of the vegetables.This paper analyse the pricing strategy by looking at the data related to each dish from 1 July to 7 July 2023, this paper fit the relationship between cost plus price and total sales using the least squares method for regression prediction, and create power function regression prediction plots for each category. The prediction charts visualise the predicted sales for the next 7 days and formulate a strategy. Through the fitting effect graph can be can be found in addition to chilli vegetables are quadratic regression, different types of vegetables sales of the annual change curve are parabolic.

Keywords: Grey Correlation Analysis, Linear Regression, Spearman Correlation Analysis.

1. Introduction

In modern agriculture and food supply chains, the study of the relationship between pricing and sales volume of vegetables is of great significance for optimizing supply chain management, formulating marketing strategies, and improving the profitability of agricultural products. With the diversification of consumer demands and the intensification of market competition, how to formulate scientific pricing strategies through data analysis and forecasting, and thus enhance the market sales volume of vegetables, has become a hot issue of concern for both academia and the industry [1].

Traditional vegetable sales analysis usually relies on empirical judgment and static data statistical methods, which make it difficult to cope with dynamic changes in the market environment [2].

Vegetable commodities in fresh produce supermarkets have a short freshness period and their quality deteriorates with the prolongation of storage time, resulting in the fact that their commodities usually cannot continue to enter the sales process on the following day after they are unsold on the same day [3]. To cope with this problem, fresh food supermarkets usually replenish their stocks daily according to the sales situation of various commodities and market demand [4]. Since supermarkets sell a wide variety of vegetables from different origins, they usually trade in vegetables between 3 and 4 a.m. during the purchase process. In this case, merchants must make replenishment decisions for the day's vegetables without knowing detailed information such as the purchase price of the specific product. Such decisions need to take into account not only the sales situation of each type of vegetable but also an accurate forecast of market demand. In terms of pricing, vegetables are usually priced on a 'cost-plus' basis [5]. However, for vegetables that have been damaged during transportation or whose quality has deteriorated, supermarkets usually choose to sell them at a discount to attract more consumers. This strategy not only helps to increase sales but also helps to reduce inventories, thereby improving overall operational efficiency. Reliable market demand analyses are key to replenishment and pricing decisions. At the demand level, there is often a correlation between the sales volume of vegetable products and the time of year. For example, consumer demand for certain vegetables may increase during certain seasons or holidays. Therefore, it is necessary to accurately predict and respond to these situations.

Firstly, the acquired data are preprocessed, and then according to the code of the single product, the acquired data are combined by the VLOOLUP function. Then the summary analysis is carried out to find the relationship between different indicators, screen out the outliers through logical operations, eliminate the outliers, and then classify and summarise the data that meets the requirements for discussion. Interrelationship analysis and grey correlation analysis of different time dimensions are carried out for vegetable categories and single-category vegetables. Finally, category summary analysis, change trend analysis, distribution frequency analysis, periodicity analysis, and single-category summary analysis were conducted.

2. Data source and preprocessing

The data for this study were obtained from www.mcm.edu.cn.

2.1. Problem analysis and data preprocessing

2.1.1. Modelling process

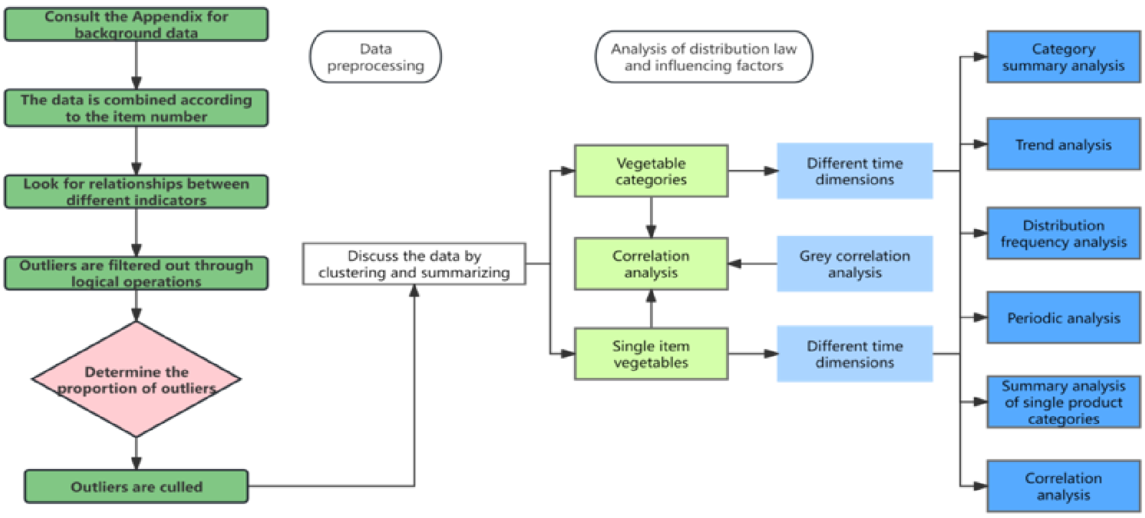

Figure 1. Modeling Process Flowchart.

The overall flow of this paper is shown in Figure 1.

2.1.2. Data pre-processing

(1) According to the single-item code, the data of single-item category data, vegetable category data, wholesale price data, and single-item loss rate are merged.

(2) Comparison of the data in Figure 2 shows that different vegetable sales flow data are missing in different months, so this paper discuss the specific situation when analyzing.

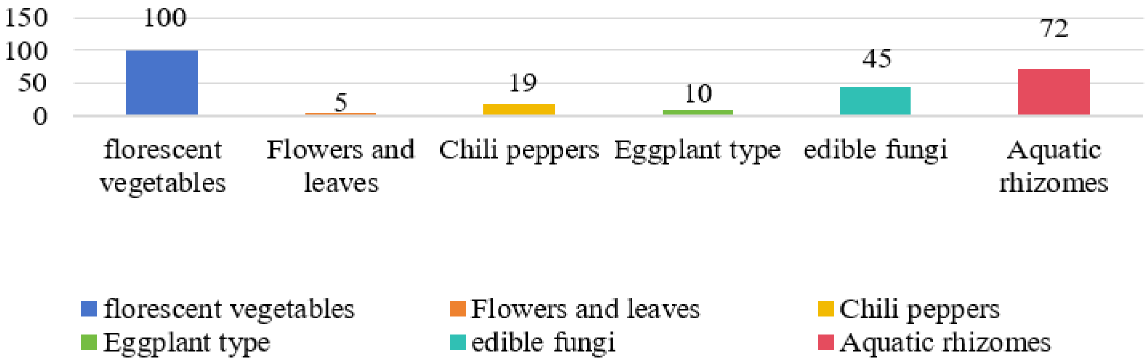

Figure 2. Classification and statistics of individual product names.

2.2. Frequency-based analysis of the distribution pattern of each category of vegetables

2.2.1. Summary analysis of vegetables by category

Aggregate analysis of the data of the six major categories of vegetables, calculated every month, yielded the aggregate trend of change over a total of 36 months from June 2020 to July 2023, as shown in the figure 3.

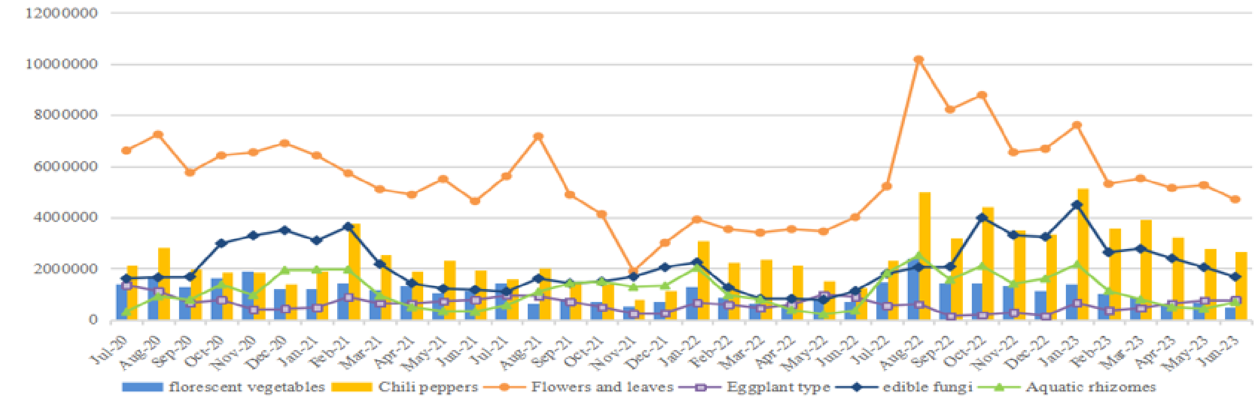

Figure 3. Summary and analysis of vegetable categories

The graph above shows that of the six main vegetables, the flower and leaf data has the highest sales and the eggplant data has the lowest sales; overall, the data fluctuates, with lower overall sales from March 2021 to June 2022, which may be affected by other factors such as the environment; after June 2022, the data shows an increasing trend; the season with the highest sales is from September to January, and the annual March to July sales are usually lower.

2.2.2. Analysis of quarterly sales trends of vegetables by category

Based on the above analysis, this paper analyzed the quarterly sales trend of the six major vegetable categories. On a month-to-month basis, this paper observed the general trend of changes in vegetable categories through linear adjustment. Based on the adjusted linear regression equation of the positive and negative coefficients, this paper determined the growth trend of change and the magnitude of change based on the absolute value of the coefficients [6]. The results are shown in Figure 4.

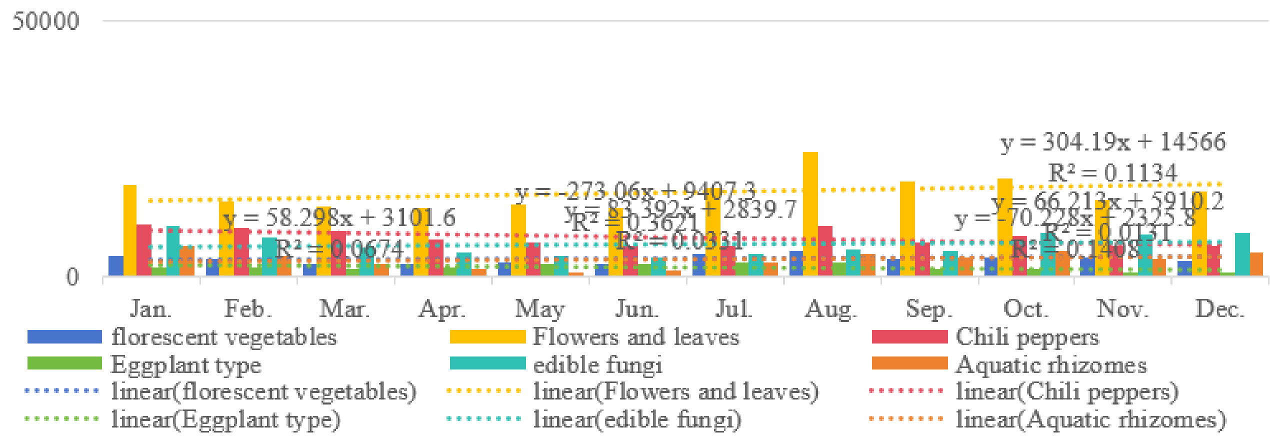

Figure 4. Analysis of Changes in Vegetable Categories.

From the figure, it is clear that cauliflower, eggplant, and foliar vegetables are showing a decreasing trend, chili and edible mushroom vegetables are showing a growing trend and aquatic root vegetables are showing a flat growth trend.

3. Vegetable Data Analysis Based on Spearman and Grey Relational Analysis

3.1. Changes in the distribution pattern of single-item sales volume based on Spearman correlation linear analysis

Firstly, the data over time is analyzed, including sample size, maximum value, minimum value, and other statistics, which are used to study the quantitative data as a whole [7]. The results are shown in Table 1:

Table1. Overall descriptive analysis of product category data.

Variable name	Sample size	maximum	average value	standard deviation	median	peak	Skewness	Coefficient
The first 18 months	246	18610.4	910.61	2491.08	6.80	20.62	4.25	2.73
After 18 months	246	15869.98	1003.92	2490.26	79.31	17.31	3.97	2.48

As can be seen from the table, the average sales of vegetables in a single category have increased over time and the median sales have increased more compared to before. At the same time, the data shows a trend of normal distribution, considering that the data did not pass the test of normal-tai distribution, so the Spearman correlation analysis is used, and the steps are as follows:

Step 1.Sort the variables x and y from smallest to largest respectively and compile the rank, which is expressed by the rank Rx and Ry. When sorting, the phenomenon of equal data resulting in the same rank becomes hold-up, in which case the average rank is taken as the rank of each data. In this case, the average rank is taken as the rank of each data.

Step 2. Spearman correlation coefficient rs is calculated by the formula:

$$r_s = \frac{\sum R_x R_y - \frac{(\sum R_x)(\sum R_y)}{n}}{\sqrt{(\sum R_x^2 - \frac{(\sum R_x)}{n})(\sum R_y^2 - \frac{(\sum R_y)^2}{n})}} \tag{1}$$

This paper analyzed the Spearman correlation coefficients of the different individual product categories in each broad category, and the results of the calculations can be concluded that the correlation between the sales of individual vegetables in the six categories of vegetables increases year by year and that the abundance of sales of vegetables in the individual category increases year by year as time changes.

3.2. Interrelationship analysis between vegetable categories and individual products based on grey correlation analysis

Grey correlation analysis is a method used to deal with multi-indicator decision-making problems with uncertainty [8]. It evaluates the degree of advantages and disadvantages of different programs or objects by calculating the degree of correlation between indicators.

For this problem, the basic process of grey correlation analysis is as follows:

Firstly, six different vegetable categories are identified as reference series, and other corresponding single categories are sub-sequences, and the data are de-quantified using Matlab to narrow down the range of variables to help simplify the calculation. Each sub-sequence collects a total of 36 samples from 1st July 2022 to 30th June 2023 in terms of months. Then our data can be expressed as.

Reference Sequence:

$$y_0 = (y_0^{(1)}, y_0^{(2)}, \ldots\ldots y_0^{(36)})^T \tag{2}$$

Subsequence:

$$x_1 = (x_1^{(1)}, x_1^{(2)}, \ldots\ldots x_1^{(36)})^T \tag{3}$$

$$x_2 = (x_2^{(1)}, x_2^{(2)}, \ldots\ldots x_2^{(36)})^T \tag{4}$$

$$x_{1091} = (x_{1091}^{(1)}, x_{1091}^{(2)}, \ldots\ldots x_{1091}^{(36)})^T \tag{5}$$

Assuming that α is the bipolar minimum difference and β is the bipolar maximum difference, there are:

$$\alpha = \min(\mathrm{i})\min(k)\left|y_0^{(k)} - x_i^{(k)}\right| \tag{6}$$

$$\beta = \max(\mathrm{i})\max(k)\left|y_0^{(k)} - x_i^{(k)}\right| \tag{7}$$

The correlation coefficient Y (Gamma value) is given by:

$$Y(y_0^{(k)}, x_i^{(k)}) = \frac{\alpha + \beta * \rho}{\left|y_0^{(k)}, x_i^{(k)}\right| + \beta * \rho} (\rho \text{ is taken as } 0.5) \tag{8}$$

Assuming a grey correlation of Y ($y_0^{(k)}$, $x_i^{(k)}$), this paper has:

$$Y(y_0^{(k)}, x_i^{(k)}) = \frac{1}{n}\sum_{k-1}^{30} Y(y_0^{(k)}, x_i^{(k)}) \tag{9}$$

The arithmetic mean root is sought for each column, which is the degree of association of each sub-sequence with the parent sequence, ultimately resulting in a flower. The grey correlation coefficients of the five sub-sequences in the Cauliflower category are visualized with the parent sequence, as shown in Figure5.

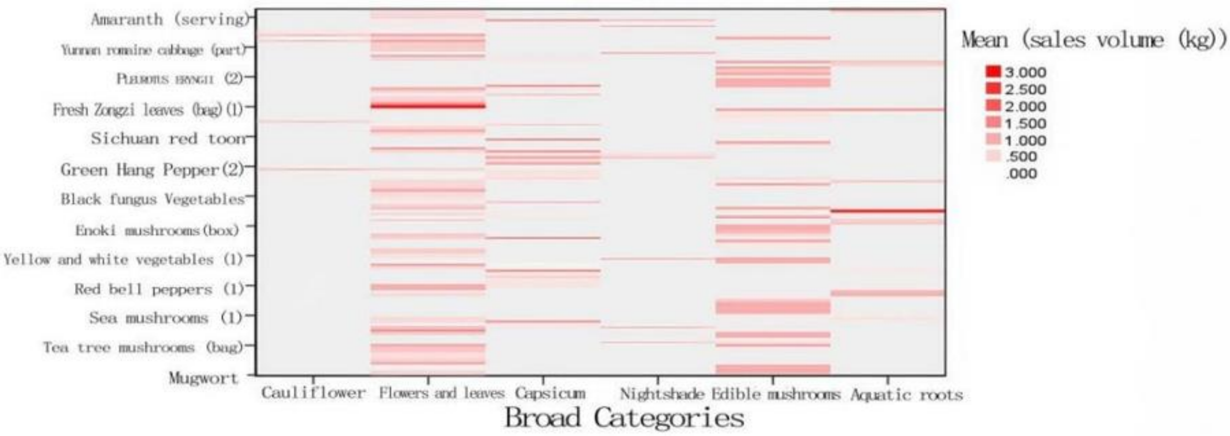

Figure 5. Visualization of the grey correlation coefficients between five single product categories and the parent sequence in the cauliflower category.

By analyzing the average value of sales of individual vegetable items as well as the broad vegetable categories by drawing heat maps, it was concluded that the sales of flowers and leaves and edible mushrooms were higher in the broad vegetable categories, while in these two categories, apricot mushrooms, fresh zongzhi leaves, and yellow cabbages also showed higher sales.

4. Cost Pricing Analysis Using ARIMA Model and Least Squares Method

After doing the distribution pattern and interrelationship of sales volume of each vegetable category and individual product to further analyze the relationship between total sales volume and cost-plus pricing of the vegetable category this is studied in depth.

Least squares regression is a method used to model and fit linear relationships and is commonly used in regression analysis [9]. Its goal is to find a straight line (or more generally a linear model) that minimizes the perpendicular distance (i.e., the sum of squares of the residuals) between the observed data points and this line, to find the line of best fit.

Below is the basic mathematical model for least squares regression:

Suppose this paper has a set of observations (x_i, y_i), where the x independent variable and y is the dependent variable. This paper wish to fit a linear model:

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{10}$$

Where y is the dependent variable, β_0 is the intercept, is the slope, and ε is the error term.

The goal of the least squares method is to find the optimal parameters β_0 β_1 such that the sum of squared residuals between the observed data points and the model predictions is minimized [10].This can be achieved by minimizing the loss function of the sum of squared residuals:

$$\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2 \tag{11}$$

Where n is the number of observed data points.

The solution of the least squares method can be obtained by solving the following two equations:

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{12}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x} \tag{13}$$

Once the best-fitting parameters β_0 β_1 and have been obtained, they can be used to make regression predictions.

Monthly changes in sales and unit prices for the broad category of vegetables are shown in Figure 6:

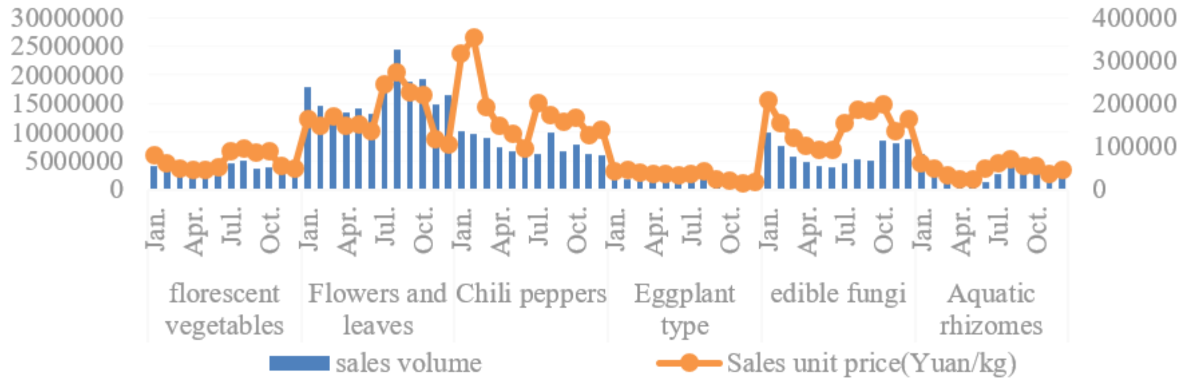

Figure 6. Monthly sales and unit price changes of vegetable categories.

The fitted scatterplot of monthly sales and unit prices for the broad category of vegetables is shown in Figure 7:

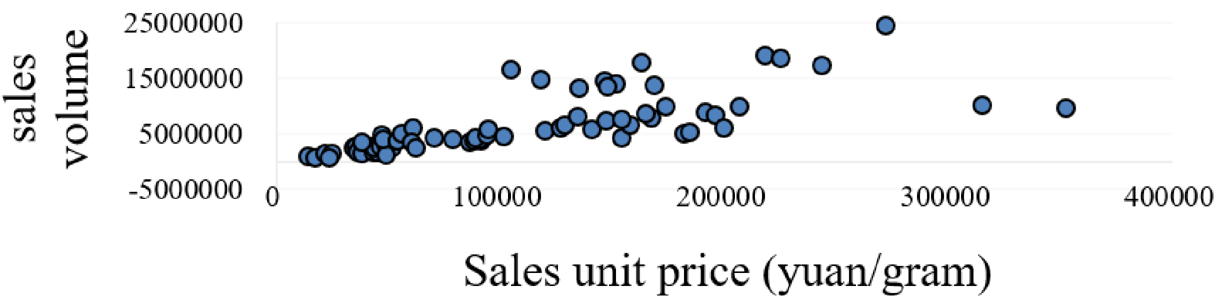

Figure 7. Scatter chart.

Analysis of the P-P plot in Figure 8 shows that the residuals basically conform to the P-P distribution and meet the regression fitting conditions.

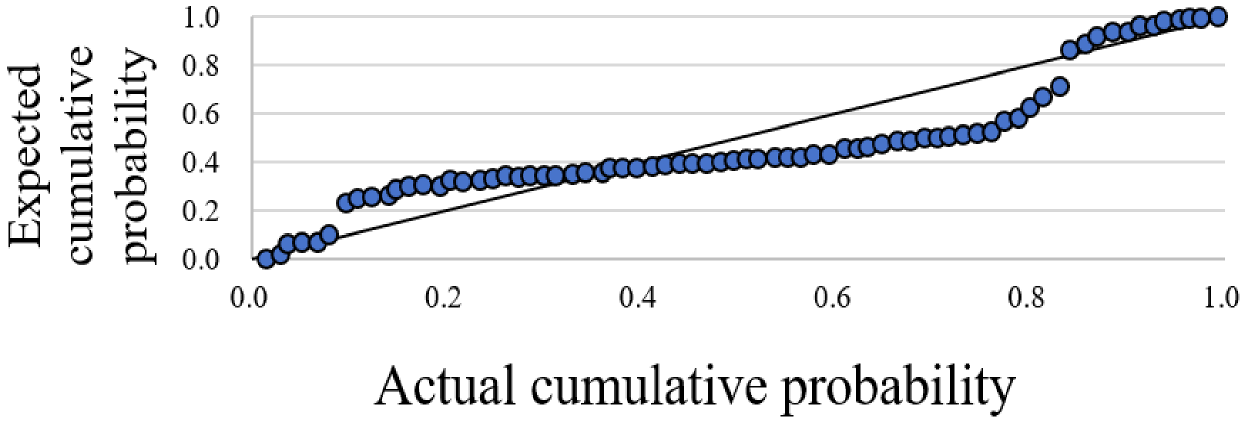

Figure 8. Normal P-P plot of standardized residuals for regression Dependent variable: Sales volume.

The predictive equations and curves are shown below by fitting the model to the unit price and volume of vegetable sales in the broad categories:

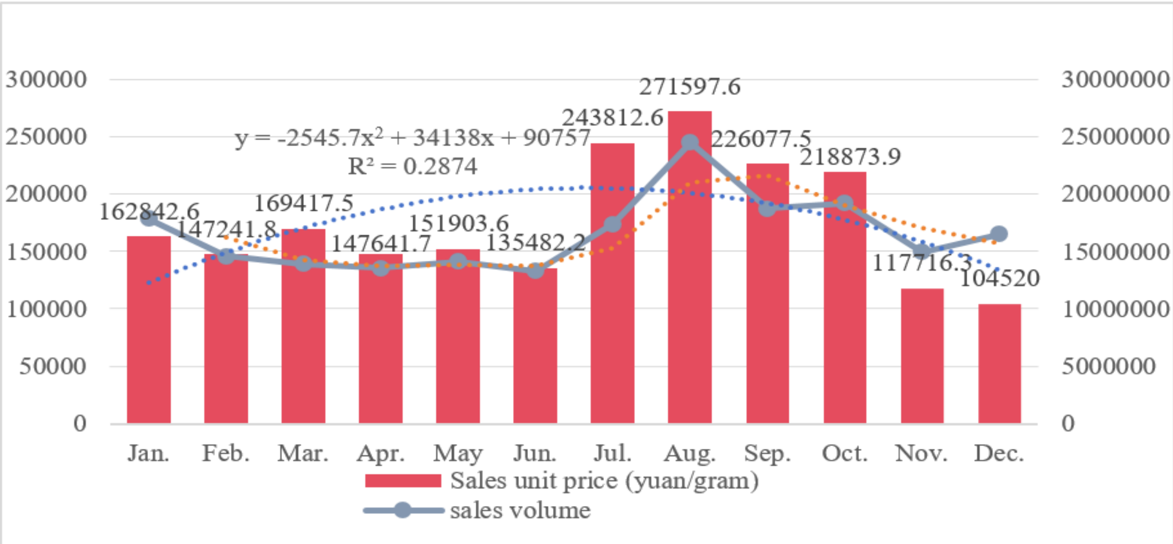

Figure 9. Flowers and leaves.

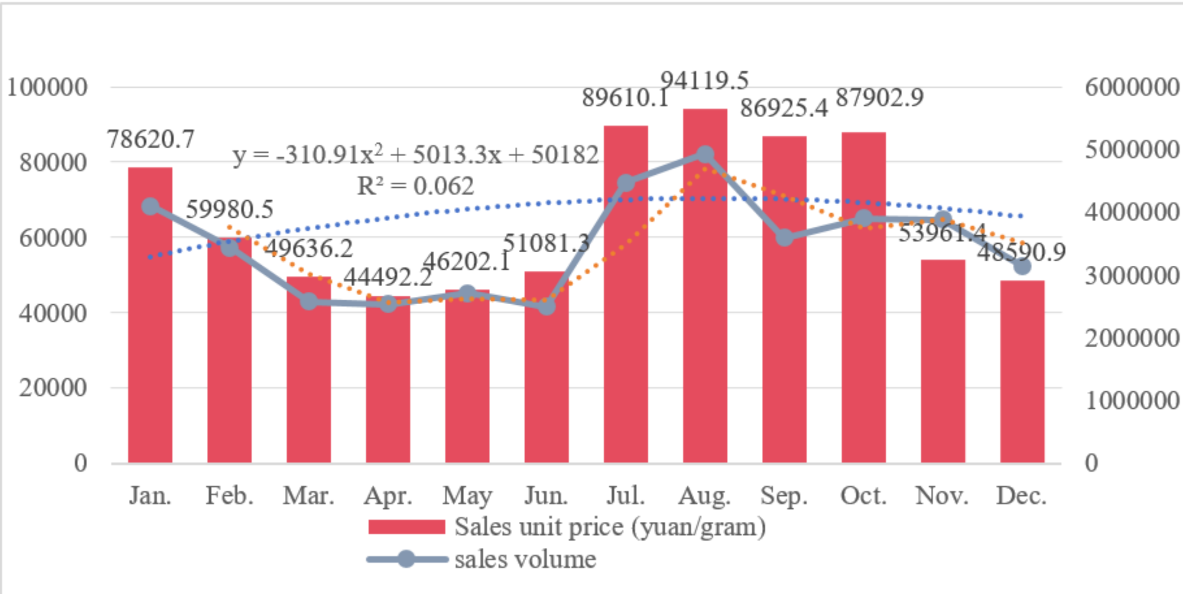

Figure 10. Florescent vegetables.

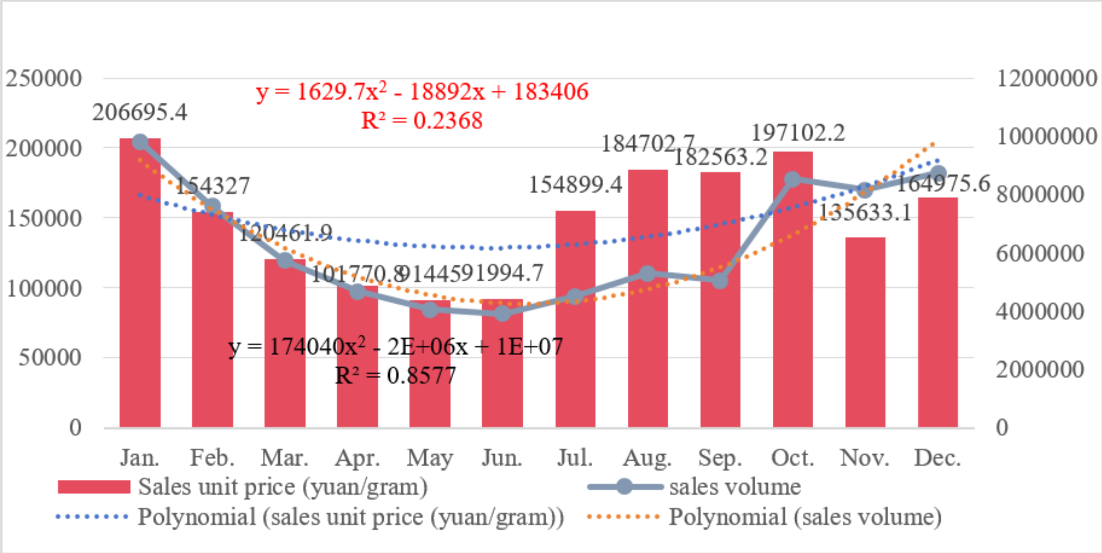

Figure 11. Edible fungi.

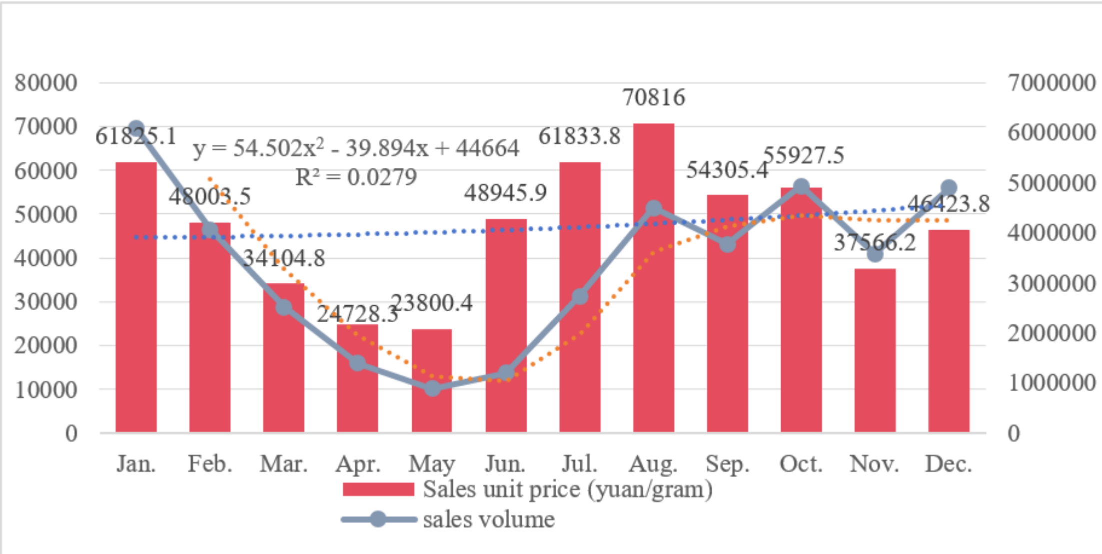

Figure 12. Aquatic rhizomes

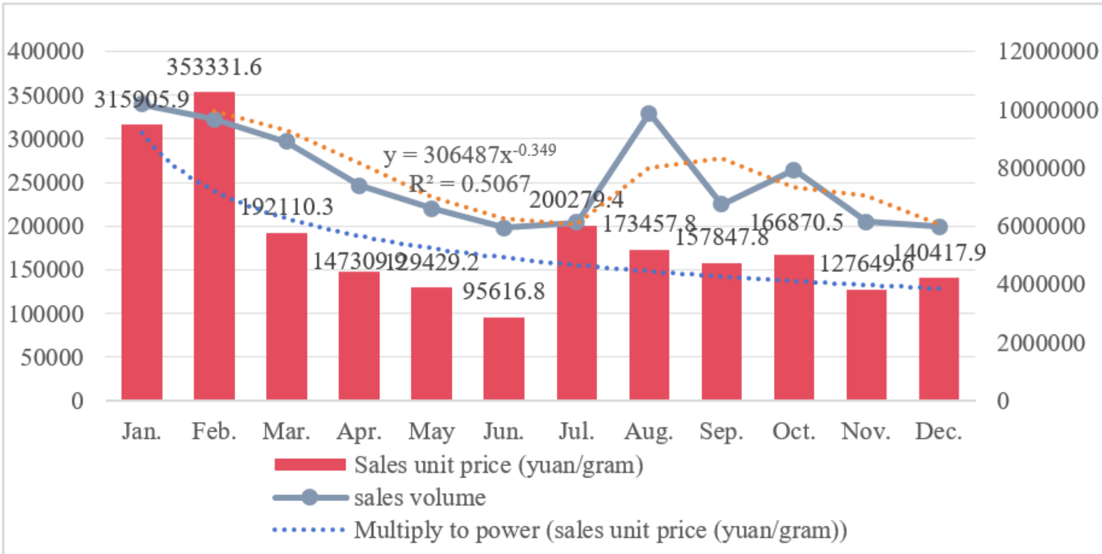

Figure 13. Chili peppers.

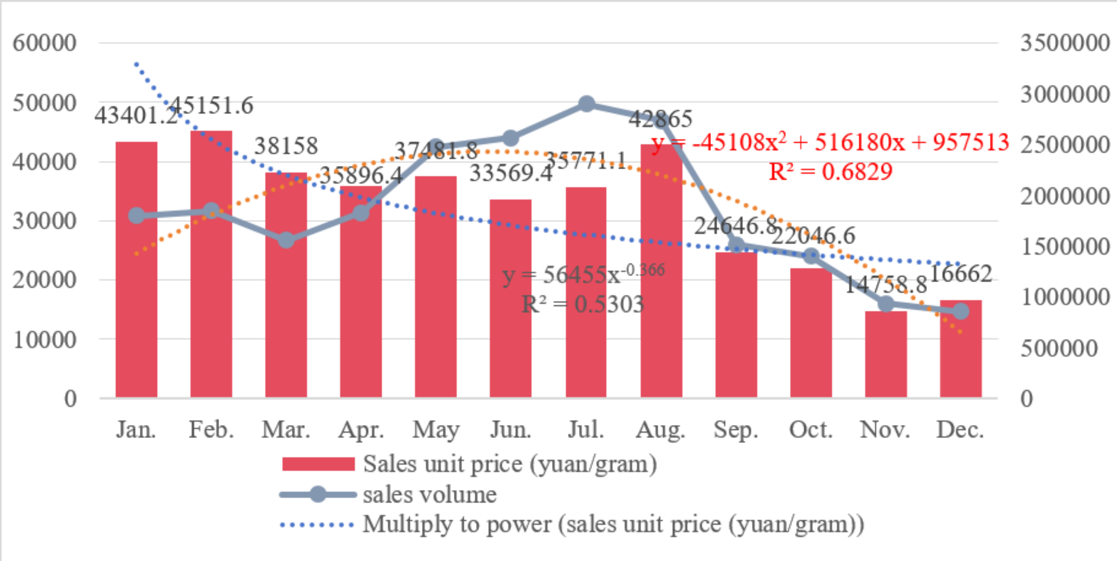

Figure 14. Eggplant type.

This paper obtained plots of the fitted effects for different categories of vegetables shown in Figures 9 to 14. The power function prediction graphs for each category visualise the predicted sales for the next 7 days. The fitted graphs show that except for the chilli vegetables, all the vegetables are quadratic regressions, and the annual change curves of the sales of the different types of vegetables are parabolic, and the non-quadratic regression of the chilli vegetables has a fitting equation of $y=306487x^{-0.349}$, where the eggplant vegetables have the highest R^2 and the aquatic root vegetables have the lowest R^2. The price of leafy and cauliflower vegetables is higher in the month of July-October as compared to other types of vegetables [11].

5. Conclusions

This study analyses the currently unclear relationships between different categories of vegetables, as well as the relationships between these categories and their respective sales volumes. Firstly, the data is visualised to analyse the distribution pattern of sales volume of each category of vegetables and individual vegetables. The results can be obtained that the daily sales volume of each single product of each vegetable category shows different degrees of seasonal fluctuations. Secondly, the Spearman correlation coefficient was introduced to the data, and the correlation analysis was carried out by taking the daily sales volume of each vegetable category and a single vegetable as the index, and the solution yielded that except for eggplant, the other five categories of vegetables showed significant positive correlations with each other. In the next step, this study explores the relationship between sales volume and cost pricing for the final pricing of vegetables.

Firstly, the daily cost-plus pricing of each vegetable category was calculated, and then the total sales volume of each vegetable category was negatively and linearly correlated with the cost-plus pricing through the Pearson's correlation coefficient test, so a linear regression model was established, and the linear regression equations were derived by using the least squares method.

The second question is to establish a time series forecasting model for different categories of vegetables, considering the timeliness of the data, the daily sales volume of the previous six months as the original data to predict the daily sales volume of each vegetable category in the coming week, and then according to the prediction results to establish the decision-making variables, profit maximization as the objective function, and restocking, pricing, and wastage of vegetables as the constraints to establish an optimization model.

Finally, the ARIMA model is invoked to find out the optimal replenishment and pricing strategies for vegetable categories.

This study provides theoretical guidance for vegetable pricing research, which not only helps retailers improve operational efficiency, reduce costs, and enhance consumer satisfaction but also supports sustainable development and promotes the development of related technologies. Therefore, this research area deserves in-depth exploration and practice.

References

[1] Wang Z, Liu X, Wang Y. Research on Vegetable Sales Volume and Pricing Strategy Based on GWO-LSTM Modeling [J]. Journal of Computing and Electronic Information Management, 2024, 12(1): 64-70.

[2] Labys W C, Pollak P K. Commodity models for forecasting and policy analysis [M]. Taylor & Francis, 2024.

[3] Mahajan P V, Caleb O J, Gil M I, et al. Quality and safety of fresh horticultural commodities: Recent advances and future perspectives[J]. Food Packaging and Shelf Life, 2017, 14: 2-11.

[4] Riesenegger L, Hübner A. Reducing food waste at retail stores—an explorative study [J]. Sustainability, 2022, 14(5): 2494.

[5] Liu J, Liu B. Commodity Pricing and Replenishment Decision Strategy Based on the Seasonal ARIMA Model [J]. Mathematics, 2023, 11(24): 4921.

[6] Alemu Z A, Dioha M O. Climate change and trend analysis of temperature: the case of Addis Ababa, Ethiopia [J]. Environmental Systems Research, 2020, 9: 1-15.

[7] Fofana F, Bazeley P, Regnault A. Applying a mixed methods design to test saturation for qualitative data in health outcomes research[J]. PloS one, 2020, 15(6): e0234898.

[8] Song Q, Wang Z, Wu T. Risk analysis and assessment of water resource carrying capacity based on weighted gray model with improved entropy weighting method in the central plains region of China [J]. Ecological Indicators, 2024, 160: 111907.

[9] James G, Witten D, Hastie T, et al. Linear regression[M]//An introduction to statistical learning: With applications in python. Cham: Springer International Publishing, 2023: 69-134.

[10] Sharma P N, Shmueli G, Sarstedt M, et al. Prediction-oriented model selection in partial least squares path modeling[J]. Decision Sciences, 2021, 52(3): 567-607.

[11] Kmeťková D, Ščasný M. Income elasticity for animal-based protein and food supply[R]. IES Working Paper, 2022.