Information Systems and Economics (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/infse.2023.041018
ISSN 2523-6407 Vol. 4 Num. 10

# *Research on Vegetable Incoming and Pricing Strategies Based on Support Vector Machines and Gray Prediction Models*

**Mengxiang Ding[1,†], Hanrui Zhang[2,†], Yiming Bao[3,†]**

[1]*School of Finance, Shandong University of Finance and Economics, Jinan, China*
[2]*School of Computer Science and Technology, Shandong University of Finance and Economics, Jinan, China*
[3]*School of Statistics and Mathematics, Shandong University of Finance and Economics, Jinan, China*
[†]*These authors also contributed equally to this work*

***Keywords:*** Market Demand Analysis, Vegetable Pricing, Mathematical Modeling, Predictive Analysis

***Abstract:*** This study analyzes vegetable stocking and pricing decisions based on market demand and historical transaction data with the aim of maximizing the profit of fresh produce supermarkets. A mathematical model was constructed using descriptive statistical analysis, cluster analysis, Pearson correlation analysis, gray prediction model, entropy weight TOPSIS method and support vector machine (SVR) regression. The study performed correlation analysis on vegetable category and specific product data and found a negative correlation between similar vegetable varieties; then it used gray prediction and TOPSIS method to predict the vegetable sales volume in the coming week and combined with SVR regression model to predict the pricing. Finally, the time series model was used to develop further pricing strategies for specific high-margin vegetable items. This study provides dedicated support for scientific decision-making on vegetable sales, as well as a reference for pricing and replenishment decisions on related items.

## 1. Introduction

In the modern fresh produce supermarket industry, limitations on the freshness period of vegetables pose a challenge to their quality and sales efficiency. The deterioration of vegetable quality over time may affect sales potential. Supermarkets face uncertainty in variety and price when purchasing diverse vegetables and need to make effective purchasing and pricing decisions. Reasonable analysis of market demand is important for developing reasonable pricing and purchasing strategies to reduce the cost of goods sold and waste while maximizing profits. The purpose of this study is to forecast the vegetable market demand, guide the future demand for vegetable categories and quantities, and provide support for replenishment and pricing decisions in fresh food supermarkets. This paper will analyze the relationship between different vegetable categories and their sales volumes, develop category-based replenishment plans and pricing strategies, and optimize

the replenishment plans and pricing strategies for individual vegetable items, considering product quantity limitations and minimum display quantities. Through these studies, this paper hopes to provide a scientific decision support system for supermarkets to improve their operational efficiency and profitability.

## 2. Relevance Analysis

### 2.1 Interrelationships among vegetable categories

In this study, Pearson correlation analysis was used to process and analyze the data with the aim of exploring statistically significant relationships between different variables. This analysis first tests whether there is a statistically significant relationship between two variables (labeled X and Y), with special attention to whether the P-value is less than 0.05 to determine its significance. Subsequently, the positive and negative directions of the correlation coefficients and their degree of correlation were analyzed. The results of correlation coefficients and model test parameters, including correlation coefficients and significance P-values, among distinct categories of vegetables were obtained through SPSSPRO software. Based on these results, this study further examined the statistical significance relationship between the two variables and analyzed the level of significance of the P-value. If the P-value shows significance (less than 0.05), it indicates that there is a correlation between the two variables; conversely, it indicates that there is no significant correlation between the two variables. This study finally summarized the positive and negative correlation coefficients and their level of correlation, which provided an in-depth analysis for understanding the interrelationships among vegetable categories [1].

Figure 1 shows the values of the correlation coefficients in the form of a heat map, which goes to indicate the magnitude of the values by the color shades.

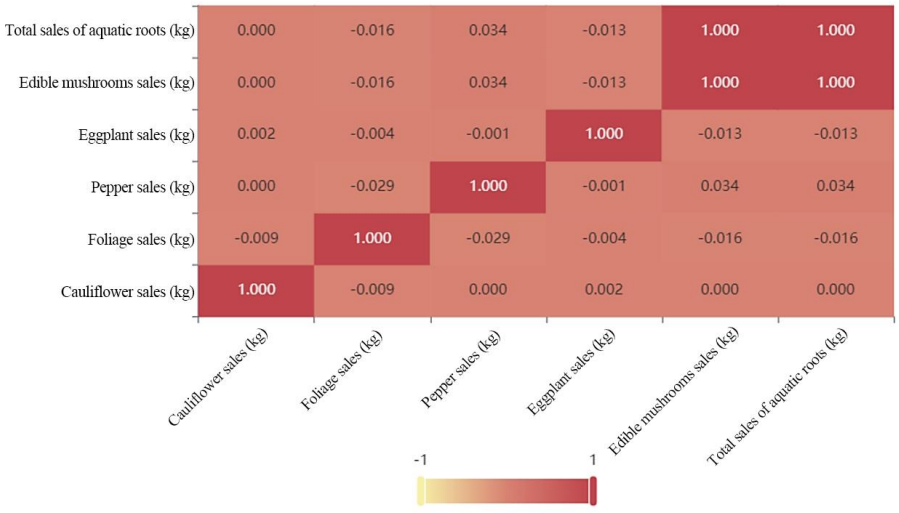


Figure 1: Heat map of correlation between distinct categories of vegetables

The histogram of sales volume of each vegetable category is plotted through MATLAB as shown in Figure 2.

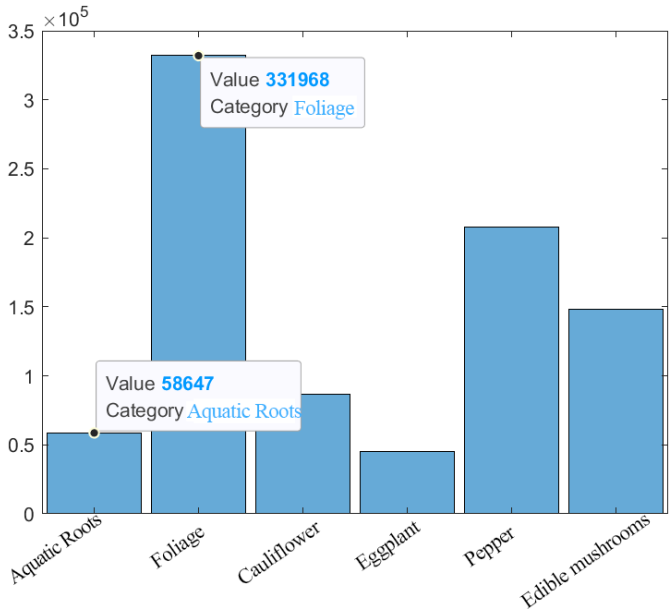


Figure 2: Histogram of Sales Volume by Vegetable Category

The correlations between different vegetable categories can be analyzed by Pearson correlation analysis:

There is a negative correlation between sales of cauliflower vegetables and sales of leafy vegetables; there is a positive correlation between sales of cauliflower vegetables and sales of eggplant vegetables; and the correlation between cauliflower vegetables and vegetables other than leafy and eggplant vegetables is weak and negligible. That is, residents' purchases of cauliflower vegetables will decrease residents' purchases of leafy vegetables like cauliflower vegetables and will increase residents' purchases of eggplant vegetables.

There is a negative correlation between the sales of leafy and flowering vegetables and cauliflower vegetables, pepper vegetables, eggplant vegetables, edible mushrooms, and aquatic root vegetables. That is, residents' purchases of flowering and leafy vegetables will affect residents' purchases of other vegetables to some extent.

There is a negative correlation between chili vegetables and leafy and eggplant vegetables, and a positive correlation between sales of chili vegetables and edible mushrooms and aquatic root vegetables. That is, residents' purchases of chili vegetables will decrease residents' purchases of leafy and eggplant vegetables and will increase residents' purchases of edible mushrooms and aquatic root vegetables.

## 2.2 Interrelationships between individual vegetable items

Based on the sales of Green Stem Scattered Flowers (kg), the coefficient of variation (CV) is 0.254, which is greater than 0.15. There may be outliers in the current data, and it is recommended to analyze the outliers or the indicators that are more prominently represented, as shown in Table 1.

Based on broccoli sales (kg), the coefficient of variation (CV) is 0.446, which is greater than 0.15. There may be outliers in the current data, and it is recommended to analyze the outliers or the indicators that are more prominently represented.

Based on the sales volume of Zijiang Green Peduncle Scattered Flowers (kg), the coefficient of variation (CV) is 0.262, which is greater than 0.15. There may be outliers in the current data, and it is recommended to analyze the outliers or the indicators that are more prominently displayed, as shown in Figure 3 and Figure 4.

Table 1: Descriptive analysis table for individual vegetable items

| Variable name | Sample size | Maximum values | Minimum value | Average value | Kurtosis | Skewness | Coefficient of variation (cv) |
|---|---|---|---|---|---|---|---|
| Sales of green stems and loose flowers | 65 | 0.865 | 0.29 | 0.487 | 0.417 | 0.571 | 0.254 |
| Broccoli sales | 65 | 1.166 | -0.436 | 0.428 | 10.137 | 0.192 | 0.446 |
| Sales of Zijiang green stem scattered flowers | 65 | 1.145 | 0.377 | 0.748 | -0.512 | 0.251 | 0.262 |
| Golden needle mushroom(box) | 65 | 1 | 1 | 1 | 0 | 0 | 0.000 |
| Almond abalone mushroom (1) | 65 | 0.992 | 0.12 | 0.31 | 7.327 | 2.159 | 0.459 |
| Fresh fungus (portions) | 65 | 1 | 1 | 1 | 0 | 0 | 0.000 |
| Apricot mushroom (2) | 65 | 0.301 | 0.164 | 0.223 | -0.141 | 0.366 | 0.133 |
| Seafood mushroom (1) | 65 | 1.101 | 0.061 | 0.238 | 18.846 | 3.594 | 0.613 |
| White mushroom(bag) | 65 | 1 | 1 | 1 | 0 | 0 | 0.000 |
| Cordyceps flowers (portions) | 65 | 1 | 1 | 1 | 0 | 0 | 0.000 |
| Red bell pepper | 65 | 0.784 | 0.031 | 0.151 | 13.512 | 2.988 | 0.771 |

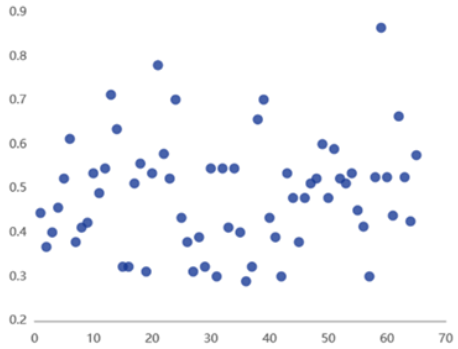


Figure 3: Scatterplot of sales volume of green stems and loose flowers

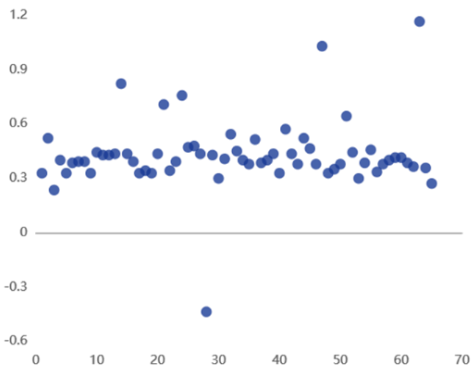


Figure 4: Scatterplot of Broccoli Sales

The above figure shows in the form of a scatterplot the sales of green peduncle loose flowers (kg), broccoli sales (kg), Zhijiang green peduncle loose flowers sales (kg), and enoki mushrooms (box). Frequency analysis concentrates on the results of trend analysis, which can be used to estimate or predict the overall, as shown in Figure 5.


Figure 5: Visualization of Pearson correlation analysis

## 3. Category pricing and replenishment strategy optimization

### 3.1 Total replenishment strategy

First of all, it is important to understand the concept of cost-plus pricing, after reviewing the information can be obtained: cost-plus price = unit into + unit cost × cost margin = unit cost × (l + cost margin). Our team calculated the cost-plus pricing for each vegetable item accordingly.

Taking cauliflower vegetables as an example, Pearson correlation analysis was applied to analyze the total sales volume and cost-plus pricing of each type of vegetables, and the results are shown in Table 2.

Table 2: Table of correlation coefficients between cauliflower sales and costs

| | Volume | Cost-plus pricing |
|---|---|---|
| Volume | 1(0.000***) | 0.94(0.687) |
| Cost-plus pricing | 0.94(0.687) | 1(0.000***) |

The following conclusions can be obtained by analyzing the above table:

There is a positive correlation between the sales volume of cauliflower vegetables and their cost-plus price, and the relationship between the sales volume of different categories of vegetables and their cost-plus price can help merchants to make decisions about the purchase volume and selling price in the coming week.

We would like to establish a gray prediction model to predict the future vegetable sales volume based on the vegetable sales volume data of previous years.

In the establishment of the gray prediction model GM (1,1) before the time series level ratio test, if the level ratio test, it means that the sequence is suitable for the construction of the gray model, if not through the level ratio test, the sequence of the "translation conversion", so that the new sequence to meet the level ratio test; gray prediction model can only be tested to determine whether it is reasonable, only through the test model can be used for prediction, the system is mainly through the a posteriori difference ratio C value to test the gray prediction model. Gray prediction models should be evaluated to determine whether it is reasonable or not. Only through the test model can be used for prediction, the system is through the a posteriori difference ratio C value to evaluate the gray

prediction model. The C-value of the posteriori difference ratio is used to evaluate the gray prediction model.

Taking cauliflower category as an example, the level ratio test is performed on the past sales volume, and all level ratio values of the level-transformed sequence are located in the interval (0.913, 1.095), which indicates that the level-transformed sequence is suitable for constructing the gray prediction model [2-3]. The parameters are shown in Table 3, and the prediction results are shown in Figure 6.

Table 3: Gray prediction model construction table

| Development factor a | Gray quantity of action b | A posteriori difference ratio C-value |
|---|---|---|
| 0.001 | 379.145 | 0.966 |

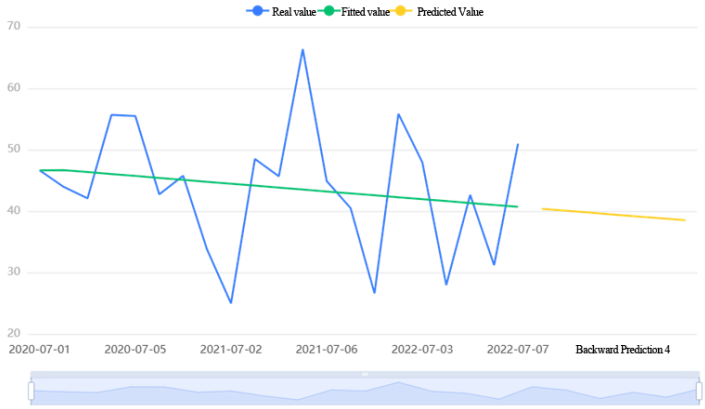


Figure 6: Model fit prediction map

The prediction of the total sales volume of other categories of vegetables is consistent with the process of predicting the total sales volume of cauliflower categories, and the results of predicting the sales volume of each category of vegetables can be obtained.

In this paper, the gray prediction model was used to predict the sales volume of each category of vegetables for the same period in 2023 based on the data of vegetable sales volume from July 1 to 7 from 2020 to 2022. In order to achieve the goal of maximizing the revenue of fresh food supermarkets, this study conducted a weighted analysis of the forecast data. The specific method is to analyze the data in depth by using the distance of superiority and inferiority solution (TOPSIS) method, which aims to adjust the weights and purchasing volume of vegetable categories. Through this method, the weights and purchases of more profitable vegetables were increased, while the weights and purchases of less profitable vegetables were appropriately reduced. This strategy helps to optimize the purchasing decision of vegetable categories, thus increasing the overall revenue of the supermarket while meeting the market demand [4]. The derived TOPSIS weighting indicators are shown in Table 4.

Table 4: TOPSIS weighting indicators

| Name | The information entropy value e | Information utility value d | Weight (%) |
|---|---|---|---|
| Total sales of flowers and foliage | 0.979 | 0.021 | 10.906 |
| Total sales of chili peppers | 0.965 | 0.035 | 18.651 |
| Total cauliflower sales | 0.97 | 0.03 | 15.595 |
| Total sales of edible mushrooms | 0.971 | 0.029 | 15.367 |
| Total sales of aquatic roots and tubers | 0.973 | 0.027 | 14.046 |
| Total eggplant sales | 0.952 | 0.048 | 25.434 |

The above table shows the results of weight calculation of entropy weight method and the weight of each indicator is analyzed according to the results. The final pricing decision for each category of vegetables was obtained by weighting the total sales volume of each category based on the weights calculated by the TOPSIS method[5].

### 3.2 Pricing strategy

Support Vector Machine (SVR) regression evaluates the model based on MSE, RMSE, MAE, MAPE, $R^2$indicators, and builds the Support Vector Machine (SVR) regression model through the training set of data; the built Support Vector Machine (SVR) regression model is applied to the training and testing data to get the model evaluation results [6].

Table 5: SVR parametric model

| Parameter name | Values of Parameters |
|---|---|
| Training time | 0.011s |
| Data Slicing | 0.7 |
| Data shuffling | No |
| Cross-validation | No |
| Penalty coefficient | 1 |
| Kernel function | linear |
| Kernel function coefficients | scale |
| Kernel function constants | 0 |
| Kernel function highest term count | 3 |
| Error convergence condition | 0.001 |
| Maximum number of iterations | 1000 |

The detailed analysis process and results are demonstrated by taking cauliflower vegetable price prediction as an example.

Table 5 demonstrates the configuration of each parameter of the model and the length of model training.

The prediction evaluation metrics of cross-validation set, training set and test set are demonstrated in Table 6, which measure the prediction effect of support vector regression through quantitative metrics. Among them, the evaluation metrics of the cross-validation set can continuously adjust the hyperparameters to obtain a reliable and stable model.

Table 6: Model evaluation results

| | MSE | RMSE | MAE | MAPE | $R^2$ |
|---|---|---|---|---|---|
| Training Sets | 2.946 | 1.717 | 1.421 | 13.501 | -0.17 |
| test set | 2.457 | 1.568 | 1.322 | 12.763 | -5.473 |

The prediction results of the test data are finally obtained. Similar to the price prediction process for cauliflower vegetables, the same can be obtained for other types of vegetables. Through the above analysis this paper derives the replenishment quantities and pricing decisions that should be made in order to maximize the profitability of the merchant, from which the maximum profitability of the daily superstore can be analyzed.

## 4. Pricing strategy optimization for individual products

### 4.1 Selected Selling Items

According to the ultimate goal of maximizing the revenue of the superstore, we conducted a statistical analysis of the profitability of each individual vegetable, based on the calculation of the profitability of each individual vegetable = (sales amount - wholesale price) × wastage rate × total number of sales to obtain the profitability of the sale of each individual vegetable, and selected 27 as the target replenishment varieties in the order of from high to low.

### 4.2 Selected Selling Items

Information on the sales of each individual vegetable item was derived from the gray prediction model, and the process here is similar to that in Section 3 and will not be repeated.

### 4.3 Developing a pricing strategy

The ARIMA model requires the series to satisfy smoothness, view the results of the ADF test and analyze whether it can significantly reject the hypothesis that the series is not smooth based on the analyzed t-value ($p<0.05$) [7].

View the data comparison graphs before and after differencing to determine whether it is smooth (not much up and down fluctuation) and bias the time series (autocorrelation analysis) to estimate its p and q values based on the truncation.

ARIMA model requires the model to have pure stochasticity, i.e., the model residuals are white noise, view the model test table, and test the model white noise based on the P-value of the Q-statistic ($P>0.05$), which can also be analyzed in combination with the information criterion AIC and BIC values (the lower, the better), and also analyzed by the model residuals ACF/PACF plots, which results in the model formula combined with the time series Analytical plots are synthesized and analyzed to obtain backward forecasts.

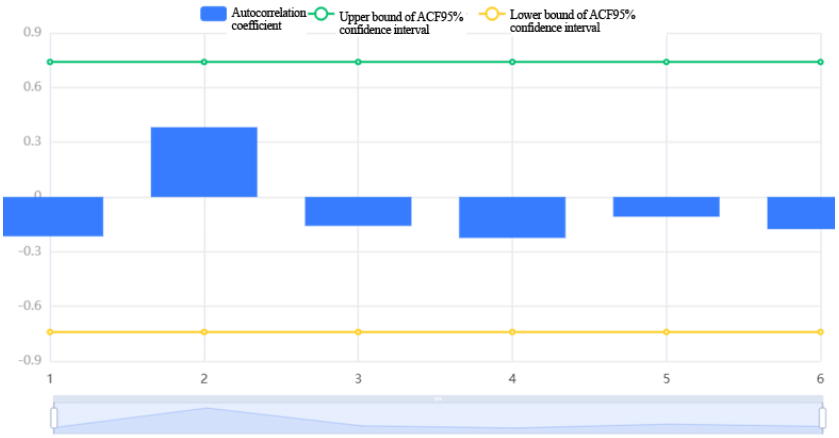


Figure 7: ACF

As shown in Figure 7, the autocorrelation plot (ACF) in this study demonstrates the autocorrelation coefficients and their upper and lower confidence limits, where the horizontal axis represents the number of delays, and the vertical axis represents the autocorrelation coefficients. The analysis shows that the ACF plot shows truncation at order q, while the partial autocorrelation (PACF) plot shows trailing, so the ARMA model can be simplified to MA(q) model. If both ACF and PACF plots are

trailing, the most significant orders in the PACF and ACF plots are used as the p and q values; if they are both truncated, a higher level of difference processing may be required or the ARMA model may not be applicable. Truncated tail means that ACF or PACF is always equal to zero after a certain order, while trailing tail means that it always has non-zero values, and these characteristics are crucial for choosing an appropriate time series model. The system automatically finds the optimal parameters based on the AIC information criterion, and the model results are ARIMA model (0,0,0) test table, based on the variable: broccoli, and from the analysis of the Q-statistic results, it can be obtained that: the Q6 does not present significance at the level, and the hypothesis that the model's residual is a white noise series cannot be rejected, while the model basically meets the requirements [8].

Finally, the optimal pricing strategy for broccoli can be obtained through the ARIMA model, and the optimal pricing results for other individual products can be obtained in the same way.

## 5. Conclusions

This paper examines the purchase and pricing decisions of vegetable-based products through descriptive statistical analysis and time-series modeling, considering the correlation between variables, and dimensionality reduction of the data through cluster analysis to simplify and objectify the results. Although data screening was performed for efficiency, which may result in slightly different results than when analyzing the entire data, and the simplicity of the model may not be applicable to complex problems, it demonstrates utility and realism in considering vegetable quality issues and circumventing the effects of epidemics. This model is not only applicable to vegetable sales, but also has implications for decision-making on other commodities with short freshness periods, helping merchants to optimize commodity allocation, reduce waste, and maximize profits.

## References


*[1] Haiqi He. Attribute approximation algorithm based on Pearson and neighborhood rough sets[J]. Information Technology and Informatization, 2023, (10):143-146.*

*[2] Duan Zhiguo, Wu Hao, Hou Yang et al. Grid cell-based point cloud dimensionality reduction algorithm[J]. Science Technology and Engineering, 2023, 23(26):11182-11187.*

*[3] Chen Pengyu, Qin Ling. Comparison of gray prediction models for deformation monitoring and study of alternative methods [J/OL]. Geodesy and Geodynamics, 1-12[2023-12-09] https://doi.org/10.14075/j.jgg.2023.07.124.*

*[4] Li P, Zhang M, Zhang HY et al. Research on multi-objective decision-making method based on rough set theory and superior-inferior solution distance method [J]. Journal of Shaanxi University of Science and Technology, 2023,41(05): 182-188.DOI:10.19481/j.cnki.issn2096-398x.2023.05.004*

*[5] Liu Q, Zhang Y, Sheng Y et al. Preferred extraction process for gastric disease 1 based on hierarchical analysis-entropy weighting-independence weighting combined with orthogonal design method [J]. China Modern Applied Pharmacy, 2023, 40(21): 2998-3004. DOI:10.13748/j.cnki.issn1007-7693.20230924*

*[6] Fan Q. Research on spatial interpolation of precipitation in Shaanxi Province based on support vector machine regression algorithm [J]. Water Resources Science and Economy, 2023, 29(09):36-40.*

*[7] Zhao Keyi. Forecast of tertiary industry development in Yunnan Province based on ARIMA model[J]. Value Engineering, 2023, 42(33):44-48.*

*[8] Wang Q. Characterization and application of spatial autocorrelation coefficient in micromotion exploration method [D]. Fujian College of Engineering, 2023. DOI:10.27865/d.cnki.gfgxy.2023.000377*