Research on Forecasting Vegetable Sales and Prices Based on Time Series Models

Xinyu Zhang*, Ruixiao Zhang, Yuan Gao

Southampton Ocean Engineering Joint Institute at HEU, Harbin Engineering University, Harbin, China, 150001

*Corresponding author: kyoutobi@hrbeu.edu.cn

Abstract. Aiming at the characteristics of fresh products that are easy to deteriorate in room temperature environment, superstores need to develop reasonable pricing strategies to improve profitability. This paper analyzes the sales data of vegetables in a superstore, uses Pearson correlation analysis and K-prototype cluster analysis to find the connection between the sales of different categories or different single vegetables, and uses the ARIMR time series forecasting model to forecast the sales and profits of different categories of vegetables. The results show that the ARIMR time-series forecasting model has good accuracy in predicting sales and profits and can provide analysis and support for pricing and replenishment decisions in supermarkets.

Keywords: Pricing, Vegetable Marketing, Time Prediction, ARIMA.

1. Introduction

Vegetable products in supermarkets usually have a short shelf life and their quality deteriorates with the increase in selling time. Most vegetables are sold at a discount on the next day if they are not sold on that day. Therefore, hypermarkets usually price and replenish daily based on the historical sales data, demand, and depletion of each commodity. Liu Hebing et al [1] used empirical modal decomposition (EMD) to decompose the price series into several intrinsic modal functions (IMFs) and cosine terms, which effectively improved the forecasting accuracy. Tang Wei et al [2] proposed a combined modeling method based on time series and multivariate analysis, which comprehensively considered the impact of historical agricultural prices and their influencing factors on forecasting. Hu Yanjun et al [3] proposed a gated logic unit neural network model incorporating the attention mechanism to make its price fluctuation characteristics more obvious. Currently, these methods are more effective in predicting individual categories of vegetables, but they do not take into account the constraints and promotional relationships that exist in the sales of different categories of vegetables. In this paper, Pearson correlation analysis and K-prototype clustering analysis are used to find the relationship between each vegetable category, and the ARIMR time series forecasting model is used to analyze the historical data of a superstore, and the validity of the model is examined. The results show that the ARIMR model can well predict the profit and sales of different categories of vegetables.

2. Correlation between category and individual product sales

Data Source: http://en.mcm.edu.cn

2.1. Analysis of sales data

This article extracts indicators from historical data to measure the multivariate elements that have a relationship with sales volume and establishes a multivariate regression model based on the resulting indicators to obtain the correlation between the sales distribution of each category and each product.

2.1.1 Exploration of vegetable sales data

The total monthly sales for different vegetable categories are visualized as shown in Figure 1.

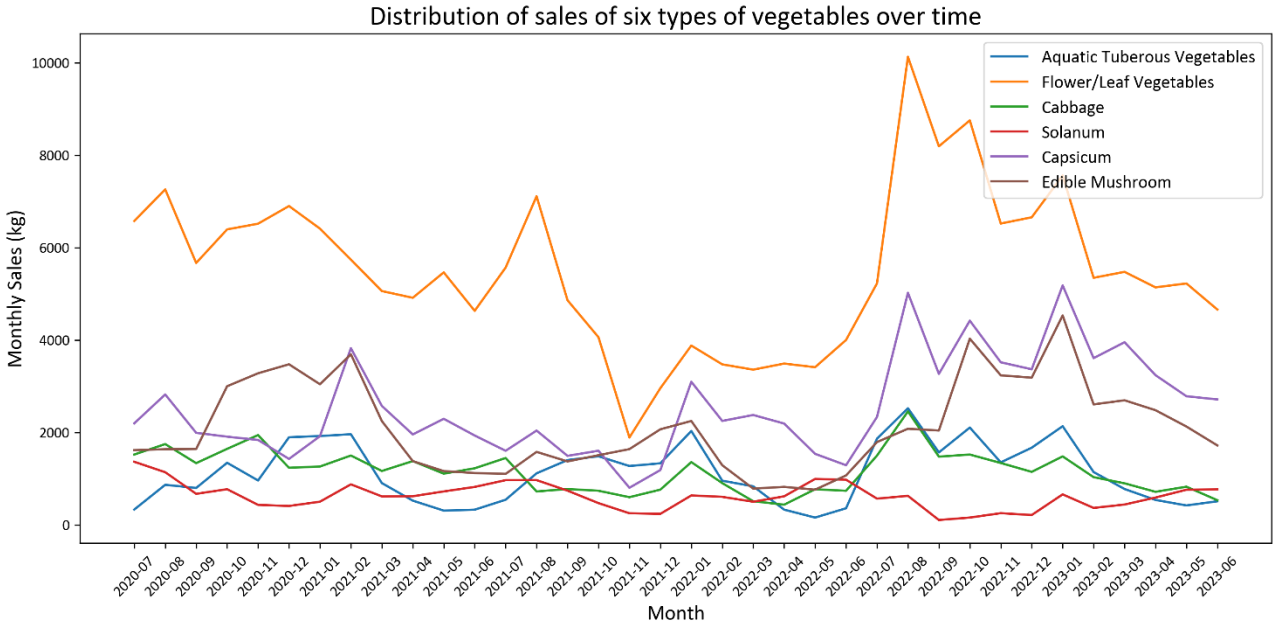


Figure 1. Distribution of sales of six types of vegetable categories

The box plot is used to represent the distribution of sales by category, as shown in Figure 2.

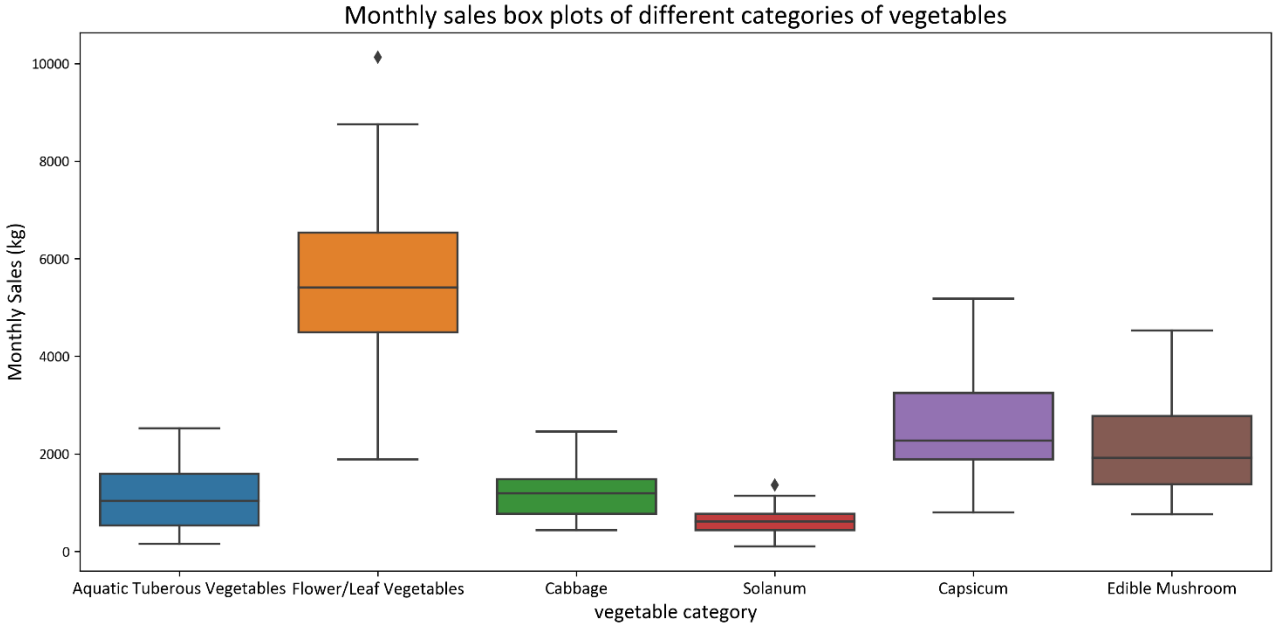


Figure 2. Monthly sales box plots of different categories of vegetables

The sampling method for individual products can reveal the existence of several single product types in the same category that are consistent, illustrating the constraints of single products in the same category and the promotion of single products in different categories.

2.1.2 Calculation of indicators

a. Standard deviation

From the standard deviation equation:

$$\sigma_Y = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - E(Y))^2}{n}} \tag{1}$$

Find the standard deviation for the total sales of each category as shown in Table.1:

Table.1 Standard deviation of total sales by category

kind	Flower/leaf vegetables	Cabbage	Solanum	Aquatic Tuberous	Edible Mushroom	Capsicum
σ_Y	86.16	22.68	13.48	31.34	48.47	53.41

b. Covariance

The overall data obtained through data preprocessing is given by the covariance formula:

$$Cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - E(x))(Y_i - E(Y))}{n} \tag{2}$$

Calculated for each of the two groups of categories, the resulting heat map is shown in Figure 3:

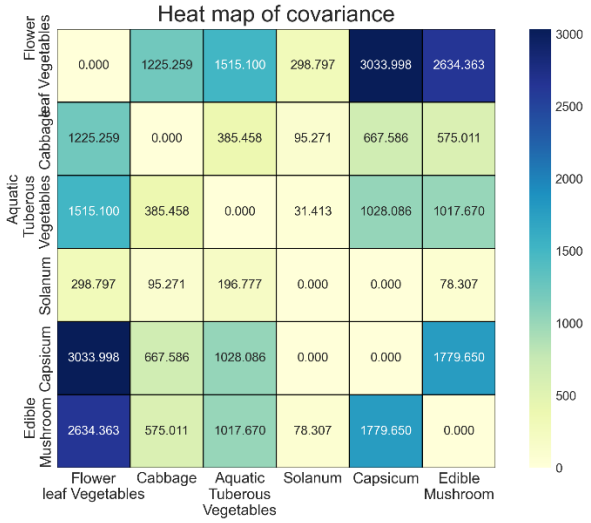


Figure 3. Heat map of covariance

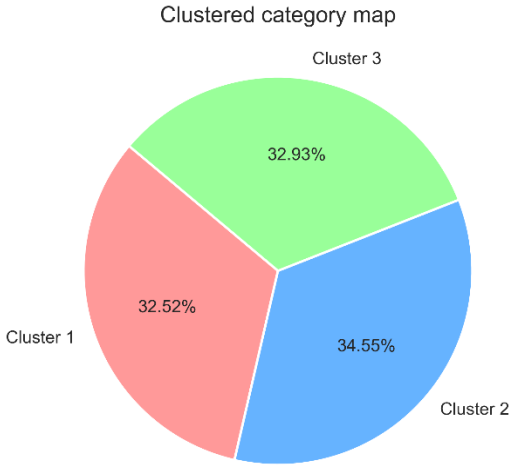


Figure 4. Clustered category map

2.1.3 Pearson model construction

After determining the indicators describing the distribution of sales, a correlation model is constructed to determine the correlation of the data through Pearson coefficient, which is a statistical indicator that expresses the degree and direction of linear correlation between two variables. The sample correlation coefficient is denoted by r and the overall correlation coefficient is denoted by ρ. Pearson correlation coefficient:

$$\rho(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_x \cdot \sigma_y} = \frac{\sum_{i=1}^{n}(X_i - E(x))(Y_i - E(Y))}{n\sigma_x \cdot \sigma_y} \tag{3}$$

2.1.4 Pearson correlation analysis for categories

(1). Flower/leaf vegetables - Cabbage, Aquatic tuberous, Edible Mushrooms, Capsicum

Pearson correlation analysis [4] was used to study flower/leaf vegetables - cabbage, aquatic tuberous, edible mushrooms, capsicum, and Pearson correlation coefficients were obtained as shown in Table.2.

Table.2. Pearson regression of flower/leaf vegetables-cabbage, aquatic tuberous, edible mushrooms, and capsicum.

kind	Cabbage	Aquatic Tuberous	Edible Mushroom	Capsicum
Flower/leaf vegetables	0.747**	0.484**	0.542**	0.624**

*p<0.05 **p<0.01

From the table, it is found that flower/leaf vegetables - cabbage, aquatic tuberous, edible mushrooms, and capsicum showed a significant positive correlation at 0.01 level of significance.

(2). Capsicum - Edible Mushrooms, Aquatic Tuberous, Cabbages

Pearson correlation analysis was used to study capsicum-edible mushrooms, aquatic tuberous, and cabbage separately and Pearson correlation coefficients were obtained as shown in Table.3.

Table.3. Pearson regression of capsicum - edible mushrooms, aquatic tuberous and cabbage

kind	Edible Mushroom	Aquatic Tuberous	Cabbage
Capsicum	0.574**	0.459**	0.425**

*p<0.05 **p<0.01

Capsicum-edible mushrooms, aquatic tuberous, and cabbage showed a significant positive correlation at 0.01 level of significance.

(3). Edible mushrooms - Cabbage, Solanum and Aquatic Tuberous

Pearson correlation analysis was used to study edible mushrooms - cabbage, Solanum, and aquatic tuberous separately and Pearson correlation coefficients were obtained as shown in Table.4.

Table.4. Pearson regression analysis of edible mushrooms - cabbage, solanum and aquatic tuberous

kind	Cabbage	Solanum	Aquatic Tuberous
Edible Mushroom	0.423*	-0.414**	0.653**

*p<0.05 **p<0.01

Edible mushrooms - cabbage and aquatic tuberous had a significant positive correlation. While edible mushrooms and solanum had a significant negative correlation.

(4) Solanum - aquatic tuberous

Pearson correlation analysis was used to study Solanum- aquatic tuberous and Pearson correlation coefficients were obtained as shown in Table.5.

Table.5. Pearson regression analysis of tomato - aquatic tuberous

kind	Aquatic Tuberous
Solanum	-0.466**

*p<0.05 **p<0.01

There was a significant negative correlation between the solanum and aquatic tuberous.

From the above table and Pearson's coefficient, we can get 21 combinations of relationships between categories, among which there are 2 groups of negative correlations, 9 groups of positive correlations, and 10 groups of no correlations.

2.1.5 Cluster analysis of individual items

A It is complex and difficult to analyze the sales volume distribution of 251 individual items using Pearson correlation analysis. In this paper, K-prototype cluster analysis [5] is used to aggregate the multidimensional data with complex structure to standardize the complex structure data. Figure 4 shows the clustering categories and the distribution of each category is uniform and good. The variability of each category group (for quantitative analysis items) was investigated using ANOVA as shown in Table.7.

Table.7. Clustered category ANOVA difference comparison results

term	difference comparison results (mean±standard deviation)			F	P
	cluster 1 (n=85)	cluster 2 (n=80)	cluster 3 (n=81)		
Product Name	43.14±24.96	206.50±23.24	125.85±23.75	954.37	0.00**
K-prototype	2.00±0.00	1.00±0.00	3.00±0.00	null	null

*p<0.05 **p<0.01

The clustered category groups showed significance (p<0.05) for all analyzed terms, indicating that the 3 category groups obtained from clustering are characterized by significant differences in Product Name. For each individual product of the same category, it is not the case that if more products are sold, more products will be bought at the same time. Despite the uncertainty of buyers' choices, it can be assumed that some of the individual products in these categories are similar when they are sold.

2.2. Sales and profit forecast

Based on the data exploration above, categories can be taken as the object of study to analyze the relationship between total sales and cost-plus pricing for each category to obtain linearly fitted regression equations, thus providing predictive data for superstores to make replenishment and pricing decisions. The relationship between sales volume and cost-plus pricing for each category is first analyzed using a logistic regression model [6]. Considering that there is no exact distribution pattern of categories per day, adopting non-linear regression method [7] can make the residual calculation more accurate.

This paper adopts a dual indicator to assess the returns, considering both sales volume and pricing rationality. Observation of historical data reveals that the sales volume of each category generally tends to increase in July, so the time series forecasting model ARIMA [8], which is applicable to the cycle, is used.

2.2.1 Logistic regression model

The modeling is based on the following equation:

$$P = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m)}{1 + \exp((\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m))} \tag{4}$$

where the model is summarized at β_0 as the constant term, and β_1, β_2, β_m as the partial regression coefficients. Performing a logit transformation on $f(x) = \frac{1}{1+e^{-x}}$, $Logit(p) = \ln\frac{p}{1-p}$,the logistic regression model can be expressed in the following linear form [9]:

$$\ln\frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m \tag{5}$$

$$\ln OR_j = \ln[\frac{p_1/(1-p_1)}{p_0/(1-p_0)}] = \ln(\frac{p_1}{1-p_1}) - \ln(\frac{p_0}{1-p_0}) = \beta_j(c_1 - c_0) \tag{6}$$

$$OR_j = \exp(\beta_j(c_1 - c_0)) \tag{7}$$

It means that the larger the OR value, the stronger the constrained correlation.

2.2.2 Logistic regression analysis

The results of ordered logistic model regression analysis showed OR = 1.046, indicating that wholesale price would have a significant positive effect on categorical names.

2.2.3 ARIMA time series forecasting model

a. Cabbage

For profit, combined with the AIC information criterion[10], several potential alternative models are modeled and compared for selection, and finally, the optimal model is identified as ARIMA(0,1,3), whose model parameter table is shown in Table.8, and the model equation is y(t)=-0.538-0.371ε(t-1)-0.309ε(t-2)-0.12ε(t-3).

Table.8. Table of parameters of ARIMA (0,1,3) model for cabbage species

term	notation	ratio	standard error	z-value	p-value	95% Cl
constant term	c	-0.538	0.933	-0.577	0.564	-2.367 ~ 1.291
	β_1	-0.371	0.028	- 13.120	0.000	-0.426 ~ -0.316
MA parameters	β_2	-0.309	0.034	-9.094	0.000	-0.376 ~ -0.242
	β_3	-0.128	0.032	-3.950	0.000	-0.192 ~ -0.065

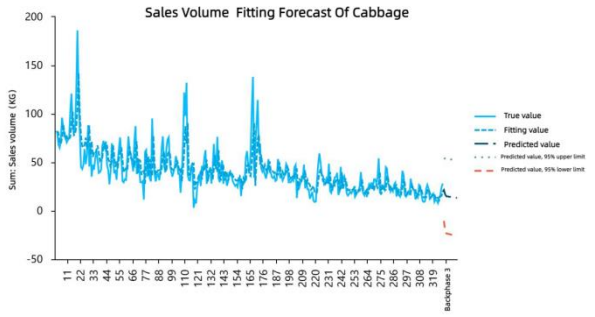
AIC value:3674.788

BIC value:3693.723

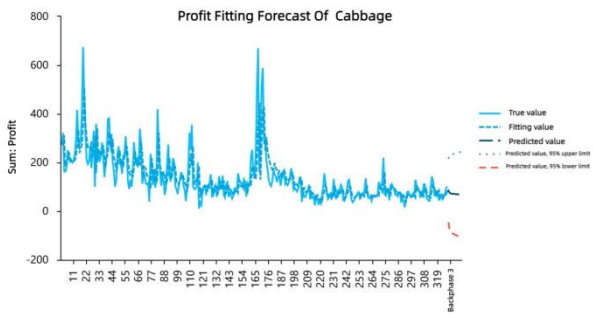
The projected values of profit and sales in the cabbage category are shown in Table.9 and the projection curves are shown in Figure 5:

Table.9 Cabbage profit and volume forecasts (12 days)

anticipate	7.1	7.2	7.3	7.4	7.5	7.6	7.7	7.8	7.9	7.10	7.11	7.12
margins	85	76	73	72	72	71	70	70	69	69	68	68
sales volume	22	17	15	15	15	15	14	14	14	14	14	13



(a) Fitted forecasts of cabbage sales



(b) Fitted forecasts for cabbage profit

Figure 5. Fitted forecasts for cabbage sales and profit

b. Capsicum

Using the same methodology as above, the optimal model was found to be ARIMA(0,1,3), whose model parameter table is shown in Table.10, and the model equation is y(t)=-0.214-0.564ε(t-1)-0.239ε(t-2)-0.145ε(t-3).

Table.10. Parameters of ARIMA (0,1,3) model for capsicum

term	notation	ratio	standard error	z-value	p-value	95% Cl
constant term	c	-0.214	0.239	-0.895	0.371	-0.681 ~ 0.254
	β_1	-0.564	0.028	- 19.967	0.000	-0.620 ~ -0.509
MA parameters	β_2	-0.239	0.044	-5.408	0.000	-0.326 ~ -0.153
	β_3	-0.145	0.040	-3.613	0.000	-0.223 ~ -0.066

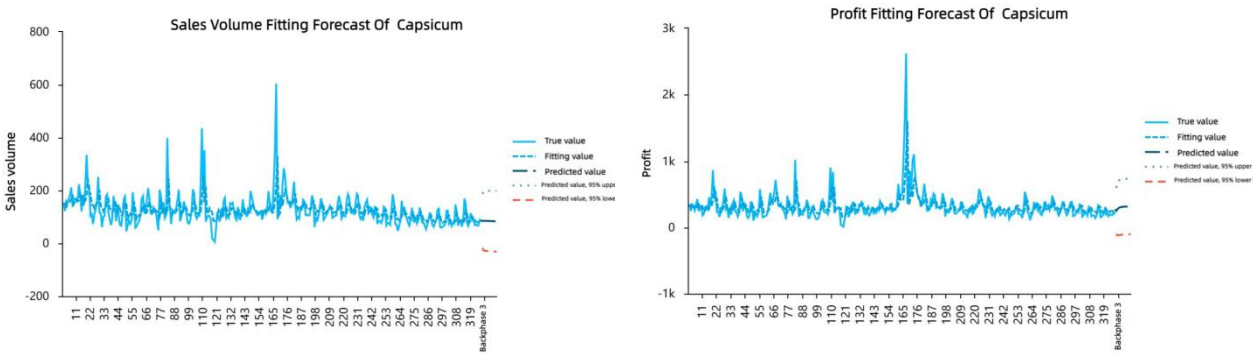
AIC value:3516.352

BIC value: 3535.286

The forecasted value of sales in the capsicum is shown in Table.11 and the forecast curves are shown in Figure 6:

Table.11. Projected value of profits and sales of capsicum (12 days)

anticipate	7.1	7.2	7.3	7.4	7.5	7.6	7.7	7.8	7.9	7.10	7.11	7.12
sales volume	85	85	85	85	85	85	85	84	84	84	84	84



(a) Fitted forecasts for capsicum sales (b) Fitted forecasts for capsicum profit

Figure 6. Fitted forecasts for capsicum sales and profit

c. Flower/Leaf Vegetables

Using the same methodology as above, the optimal model was found to be ARIMA(3,1,1), whose model parameter table is shown in Table.12, and the model equation is y(t)=-0.356+0.218y (t-1)+0.254y (t-2)-0.119y (t-3)-1.000ε (t-1).

Table.12. Parameters of the ARIMA (3, 1, 1) model for flower/leaf vegetables classes

term	notation	ratio	standard error	z-value	p-value	95% Cl
constant term	c	-0.356	0.107	-3.328	0.001	-0.566 ~ -0.146
AR parameters	α_1	0.218	0.022	9.947	0.000	0.175 ~ 0.260
	α_2	0.254	0.024	10.390	0.000	0.206~ 0.302
	α_3	-0.119	0.076	-1.567	0.117	-0.268 ~ 0.030
MA parameters	β_1	-1.000	2.622	-0.381	0.703	-6.140 ~ 4.140

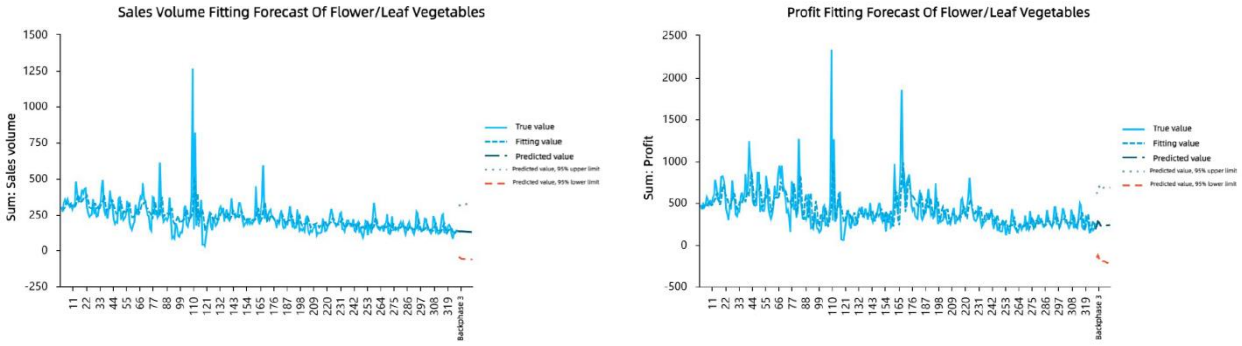
AIC value:3882.340

BIC value:3905.061

The projected value of sales in the flower/leaf vegetables is shown in Table.13 and the projection curves are shown in Figure 7:

Table.13. Flower and Leaf Profit and Volume Forecasts (12 days)

anticipate	7.1	7.2	7.3	7.4	7.5	7.6	7.7	7.8	7.9	7.10	7.11	7.12
margins	226	277	264	230	222	237	246	240	232	231	234	235
sales volume	134	133	134	133	133	132	132	131	131	130	130	129



(a) Fitted forecasts for flower/leaf vegetables sales

(b) Fitted forecasts for flower/leaf vegetables profit

Figure 7. Fitted forecasts for flower/leaf vegetables sales and profit

3. Conclusion

For the huge commodity data, this paper first used Pearson correlation analysis and K-prototype cluster analysis methods to process the data with different characteristics and obtain the correlation relationship between the sales volume of each category and each product. Logistic regression model was used to analyze the correlation relationship between the sales volume of each category and the cost-plus pricing, and then the sales volume and profit of each category of vegetables in July 2023 were predicted by the ARIMA time series forecasting model. To verify the feasibility of the model, this paper compares the real values of cabbage, capsicum, and flower/leaf vegetables with the predicted value curves, and the results show that the prediction accuracy is good. The model is suitable for future heartbeat decision analysis based on historical data, including but not limited to

buying and selling pricing, hiring short-term workers, buying and selling stocks, and other predictions through past data can also be solved by this model.

References

[1] Liu Hebing, Han Jingjing, Zhuang Chenhui et al. Research on vegetable price prediction model based on multi-scale feature fusion [J]. Journal of Henan Agricultural University, 2022, 56 (05): 858-867.

[2] Tang W. Combined modeling-based price forecasting of agricultural products [D]. Sichuan Agricultural University, 2023.

[3] Hu Yanjun, Zhang Pingchuan, Shang Zheng, et al. Research on garlic price prediction based on deep learning [J]. Journal of Henan Institute of Science and Technology (Natural Science Edition), 2023, 51 (03): 35-42.

[4] WANG Xinyuan,LI Jian. Analysis of operational faults and risk control countermeasures of Shanghai metro based on K-prototypes clustering algorithm[C]// Academic Committee on Urban Transportation Planning, China Society of Urban Planning. Green-Wisdom-Integration-Proceedings of the 2021/2022 Annual Conference on Urban Transportation Planning in China. School of Transportation Engineering, Tongji University;, 2022: 11.

[5] Hu XH, Ye JL, Sheng JX. Precision marketing strategy for e-commerce in big data environment based on K-Means cluster analysis [J]. Journal of Langfang Normal College (Natural Science Edition), 2023, 23 (04): 50-52+79.

[6] LIU Entertainment,TIAN Maozai. Research on APP task pricing application based on logistic regression model [J]. China Price, 2020, (08): 98-101.

[7] HU Wenhui,CHEN Shunxuan,FU Junkai. Task pricing law model based on nonlinear regression [J]. China Hi-Tech Zone, 2018, (07): 29.

[8] Mao Lisha. Research on pricing strategy and production and marketing model of vegetable wholesale market under the perspective of supply chain [D]. Central South Forestry University of Science and Technology, 2022.

[9] ZHAO Tianwei, CHEN Huida. National population forecast based on linear regression model, Malthusian population growth model and logistic model [J]. Journal of Guangdong Medical University, 2023, 41 (06): 623-627.

[10] Xie Hai-Ning,Shang Yuan. Stock price prediction based on ARIMA modeling analysis--Taking CITIC Bank as an example [J]. North Economy and Trade, 2020, (12): 122-124.