Transactions on Computational and Applied Mathematics (2024)
Clausius Scientific Press, Canada
DOI: 10.23977/tracam.2024.040103
ISSN 2616-1826 Vol. 4 Num. 1

Automatic Pricing and Replenishment Decision Making for Vegetable Products Based on SARIMA and Nonlinear Programming

Gaofeng Zhu[1,#], Luyao Xin[2,#], Wenhuan Yang[2,#]

[1]*College of Enginerring, Huaqiao University, Quanzhou, 362021, China*
[2]*School of Mathematical Sciences, Huaqiao University, Quanzhou, 362021, China*
[#]*These authors contributed equally.*

Keywords: Pricing and Replenishment of Vegetable Commodities, Cosine Similarity, SARIMA, Multiple Logistic Regression, Nonlinear Programming

Abstract: In this paper, for the automatic pricing and replenishment decision problem of vegetable goods, firstly, vegetable single items are divided into four categories according to sales volume, and the statistics of the number of vegetable single items with newly classified vegetable single items in each category are counted, and the cosine similarity is calculated, and it is found that: the similarity between the various categories is extremely high, and it is generally higher than 0.8. Then, three vegetable single items of Niushou lettuce, broccoli and net lotus root (1) are randomly selected, and the selected items are subjected to a smooth. Then, three vegetable items were randomly selected, and a smooth time series test was performed on the selected items, and the significance P-values of 0.000, 0.001 and 0.032 were obtained, indicating that the distribution of sales volume of vegetable items showed significant seasonality. Multivariate logistic regression was then used to show that total sales volume is negatively related to cost-plus pricing. With regard to the total daily replenishment and pricing strategy of each vegetable category in the coming week, the objective function is constructed with the orientation of maximizing the supermarket's revenue, and the maximum value of the total daily replenishment in the coming week is predicted by using SARIMA, and then considering the quantity constraints of each category and other conditions, the maximum value of the product's revenue from 1 to 7 July 2023 is solved to be ¥15,214.83, and the corresponding daily replenishment and pricing strategy is obtained. The total daily replenishment and pricing strategy are obtained accordingly.

1. Introduction

Agriculture plays an important role in China's economic development, and the vegetable market occupies a certain position in the agricultural economy. In the vegetable sales process, the quality of vegetables usually deteriorates with the increase of sales time, and the pricing of vegetables is generally based on the "cost-plus pricing" method, which requires supermarkets to make replenishment and pricing decisions on a daily basis according to the sales and demand situation.

(1) Factors affecting vegetable prices:

Boqiao Yuan (2023) concluded that short-term fluctuations in vegetable prices have seasonal characteristics by setting up the VECH model, and the transmission mechanism of vegetable price fluctuations in the short term may be affected by its own pricing mechanism[1-2] .

(2) Research on pricing issues related to agricultural products:

Yajie Lu (2010), by analyzing the pricing trend of modern retail commodities, proposed a new method of price adjustment for high-quality fresh vegetables operated by supermarkets-value loss pricing method [3]; Sihong Gu (2023), based on the background of the fresh market and the status quo of the retailers, found out that the fresh products due to the decline in freshness of the products and slightly too high the problem of profit loss [4];

(3) Research methods related to agricultural pricing issues:

Zhang Lu (2019) proposed clustering variables by maximum distance correlation coefficient method [5]; Jifu Qian (2010) constructed a new suitable method for modeling seasonal time series forecasting based on SVR, which can show better forecasting performance[6].

To summarize, vegetable prices are affected by many influences in the supply chain process, while this paper investigates the similarity between each vegetable category, the relationship between the sales volume of a single vegetable item and the season, and the relationship between the total sales volume of each vegetable category and the cost of pricing; this paper uses data analysis methods such as cosine similarity analysis, SARIMA, and multivariate Logistic regression in its research methodology.

2. Data exploration on vegetable sales volume

Data from http://www.mcm.edu.cn/ data include individual item code, individual item name, classification code, classification name, date of sale, time of sweeping and sale, sales volume (kg), unit price of sale (¥/kg), type of sale, date of wholesaling, and wholesaling price (¥/kg), Loss rate (%), etc.

2.1 Distribution patterns and interrelationships among vegetable categories

In this paper, the total sales volume of 244 vegetable items from July 1, 2020 to June 30, 2023 was calculated using Python, and the total sales volume of the remaining 7 vegetable items was zero.

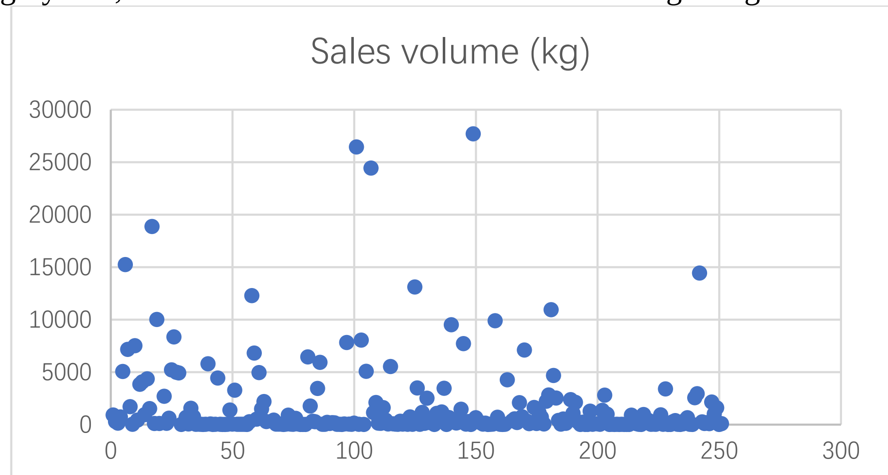


Figure 1: Distribution of total sales of individual vegetable items

In order to study the distribution pattern of each category of vegetables, we decided to use a visualization tool to initially explore the distribution of the total sales volume of each vegetable item, as shown in Figure 1:

According to Figure 2, we can categorize the 251 vegetable items into 4 categories according to the total sales volume, and the total sales volume of each category is 0, (0, 1000), (1000, 10000), (10000, 30000). The classification results are shown in Table 1:

Table 1: Vegetable single product sales range no.

Vegetable single product sales range	Classification Number
0	1
(0, 1000)	2
(1000, 10000)	3
(10,000, 30,000)	4

In this paper, cosine similarity is used as a similarity metric to calculate the interrelationships between vegetable categories. Cosine similarity measures the degree of similarity between two vectors. By calculating the cosine similarity between each vegetable category, we can derive their similarity profile.

Cosine similarity is a metric used to measure the degree of similarity between two vectors, and is commonly used in areas such as calculating text similarity and recommender systems. The principle is to measure the similarity between two vectors based on the angle of the vectors. The smaller the angle, the more similar the two vectors are; the larger the angle, the less similar the two vectors are.

The cosine similarity formula is as follows:

$$\cos(A, B) = \frac{A \cdot B}{|A||B|}) \quad (1)$$

where $A \cdot B$ denotes the inner product of vector A and vector B, and $|A|$ and $|B|$ denote the modulus of vector A and vector B, respectively.

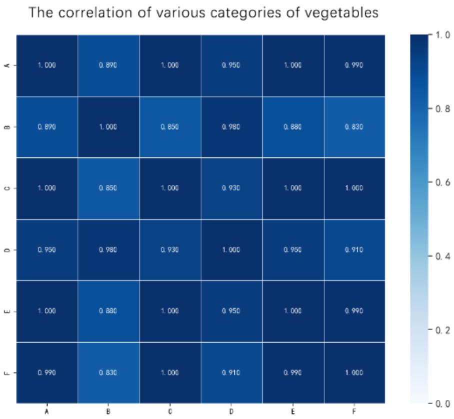
With the above steps, the cosine similarity between the two vegetable categories can be calculated and used to measure how similar they are. The value of cosine similarity ranges from -1 to 1. The closer the value is to 1, the higher the similarity, and the closer the value is to -1, the lower the similarity.

Considering the problem of uneven distribution of the number of vegetables in each vegetable category, in order to avoid the impact on the similarity analysis, we denote the distribution of single products as the vector A_i (i=1,2,...6), whose data in each dimension are in turn a_{ij} (j =1,2,...4). a_{ij} denotes the number of i vegetable category has the first j The ratio of the number of vegetable items in the category to the number of vegetables in the category. Take vegetable category B as an example, it owns 0 vegetable items of the first category, 2 vegetable items of the second category, 2 vegetable items of the third category, and 1 vegetable item of the fourth category. $A_2 = [0,40,40,20]$, Similarly, can be calculated $A_1, A_3, \ldots A_6$, and the results are shown in Table 2:

Table 2: Distribution of individual items by vegetable category

vegetable category	Category 1	Category 2	Category 3	Category 4
Foliage	3.00	66.00	27.00	4.00
Cauliflower	0.00	40.00	40.00	20.00
Aquatic Roots	5.26	68.42	21.05	5.26
Eggplant	0.00	50.00	40.00	10.00
Pepper	4.44	66.67	26.67	2.22
Edible Mushroom	1.39	75.00	20.83	2.78

Next, for the distribution of selected individual items among different vegetable categories A_i Cosine similarity is calculated and heat map is plotted for the next analysis as shown in Fig. 2:



Note: A, B, C, D, E, F stand for Foliage, Cauliflower, Aquatic Roots, Eggplant, Pepper, and Edible Mushroom, respectively.

Figure 2: Similarity in the distribution of individual items in each category of vegetables

Based on the similarity heat map of the distribution of single items in each category of vegetables we can get the following conclusions:

(1)The distribution of individual items among the vegetable categories is highly correlated.

(2)The pattern of distribution of individual items between the three categories of foliage, aquatic roots and peppers was almost identical.

(3)The similarity in the distribution pattern of individual items between the two categories of cauliflower and pepper is relatively low.

2.2 Distribution pattern of sales volume of individual vegetable products

SARIMA is a linear method capable of modeling many seasonal time series. Smoothness is the basis of time series analysis; a time series can be smooth or non-smooth. A smooth time series is more accurate in predicting future changes in trend, therefore, in time series analysis, usually the series must be smooth. The following three conditions must be met for a time series to be smooth:

$$E(Y_t) = \mu \tag{2}$$

$$Var(Y_t) = \sigma^2 \tag{3}$$

$$Cov(Y_t, Y_{t+k}) = \gamma_k \tag{4}$$

Of these, theμ andσ is a constant independent of time;γ_k is a constant related only to the time interval k and independent of t.

If a time series is non-stationary, it has both the characteristics of a random walk, and the random walk can be expressed as follows:

$$Y_t = Y_{t-1} + \alpha_t \tag{5}$$

$$Y_t = \mu + Y_{t-1} + \alpha_t \tag{6}$$

where $\{\alpha_t\}$ is a white noise sequence, Eq. (2) inμ is a constant term, and Eq. (2) represents a random wandering with drift.

Random wandering generally can not predict the changes in the next period, so the non-smooth time series must first be smoothed, the commonly used method of treatment is differential; if for seasonal time series, generally must also be seasonal differential to eliminate seasonal trends, so that the sequence to achieve smooth.

To determine whether a series is smooth can be roughly judged by looking at the time path of the series graph: a smooth time series in the graph tends to show a fluctuation around the mean value of the process; and non-smooth series tend to show in different time periods have different mean values. It can also be determined by examining the sample autocorrelation function and its graph. The autocorrelation function (ACF) is defined as follows:

$$\rho_k = \frac{\gamma_k}{\gamma_0} \tag{7}$$

where the numerator is the covariance of the series lagged by k periods and the denominator is the variance. The autocorrelation function is a decreasing function with respect to lag k. Since in practice we have only one realization sample for a stochastic process, only the sample autocorrelation function can be computed and the sample autocorrelation function is defined as follows.

$$\gamma_k = \frac{\sum_{i=1}^{n-k}(Y_t-\bar{Y})(Y_t-\bar{Y})}{\sum_{i=1}^{n}(Y_t-\bar{Y})^2} \tag{8}$$

As k increases, the sample autocorrelation function decreases and tends to 0. However, in terms of the rate of decrease, the smooth series is much faster than the non-smooth series, which is a way to determine whether the series is smooth or not.

In general, for a system consisting of SARIMA $(p, d, q\ )(P, D, Q)_s$ process produces a sequence $\{Y_t\}$, the model can be formulated as follows using the backward push operator:

$$\phi_p(B)\Phi_P(B^s)(1-B)^d(1-B^s)^D Y_t = \theta_q(B)\Theta_Q(B^s)\varepsilon_t \tag{9}$$

where B is the backward pushback operator (definition of$B^K Y_t = Y_{t-k}$);

$$\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p \tag{10}$$

$$\Phi_p(B^s) = 1 - \Phi_s B^s - \Phi_{2s} B^{2s} - \cdots - \Phi_{ps} B^{ps} \tag{11}$$

$$\theta_p(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q \tag{12}$$

$$\Theta_Q(B^s) = 1 - \Theta_s B^s - \Theta_{2s} B^{2s} - \cdots - \Theta_{Qs} B^{Qs} \tag{13}$$

p, d, q, P, D, Q are non-negative integers, and d is the number of regular differentials, and D is the number of seasonal differentials, and p is the order of autoregression (AR), and$P$ is the order of seasonal autoregression (SAR), and$g$ is the order of the moving average (MA), and O is the order of the seasonal moving average (SMA). s denotes the seasonal length, with 4 denoting the seasonal series and 12 denoting the monthly series E, a white noise residual series with zero mean and constant variance.

In order to study the distribution pattern of vegetable singles, we randomly selected three vegetable singles as research objects, namely, beef head lettuce, broccoli, and net lotus root (1), and used SARIMA to explore the seasonal cyclic pattern of the above three singles, and due to the space limitation, this paper only demonstrates the analysis process of beef head lettuce.

Based on the sales of individual vegetable items from 2020-7-1 to 2023-6-20 provided by the raw data, the monthly sales of the three vegetables can be calculated on a month-by-month basis as a way of exploring the distribution pattern of sales versus time.

The SARIMA model was first used to explore the pattern of the distribution of Oxalis lettuce with respect to time, as shown in Table 3:

Table 3: Table of results of smooth time series test for beefsteak lettuce

variant	sequences	t	P	AIC	threshold value		
					1%	5%	10%
series1	primary sequence	-1.408	0.578	66.036	-3.689	-2.972	-2.625
	1st order difference	-5.103	0.000***	257.072	-3.689	-2.972	-2.625
	1st order difference - 1st order seasonal difference	-4.738	0.000***	216.204	-3.77	-3.005	-2.643
	2nd order difference	-5.2	0.000***	255.39	-3.7	-2.976	-2.628
	2nd order difference - 1st order seasonal difference	-3.936	0.002***	192.001	-3.889	-3.054	-2.667
Note: ***, **, * represent 1%, 5%, and 10% significance levels, respectively.							

The result table of this sequence test shows that the 1st order difference sequence based on beefsteak lettuce, with a significance p-value of 0.000***, presents significance at the level, rejecting the original hypothesis that the sequence is a smooth time series, which can be explored using the selected seasonal time series, and obtaining the original, trend, and seasonal sequences as shown in Figure 3:

Figure 3: Graphical representation of seasonal time series correlation of lettuce sales in beef heads

From the original sequence diagram, the trend sequence diagram, and the seasonal sequence diagram, it can be seen that (1) Beef head lettuce shows a very significant upward trend in sales in November each year until it reaches the highest value of sales for the year. (2) Beef head lettuce shows a significant downward trend from April to May each year, and reaches the lowest sales value of the year from July to November. (3) The sales of beefsteak lettuce show a more obvious seasonality, with a significant increase in sales in winter and a significant decrease in summer and fall. (4) The monthly sales volume of lettuce was low, with a minimum of 0 kg and a maximum of less than 300 kg per month. Broccoli and lotus root (1) were analyzed in the same way as lettuce.

3. Pricing and replenishment decision-making for vegetables

3.1 Relationship between total sales and cost-plus pricing in the vegetable category

The mathematical expression for the Sigmoid function is:

$$f(x) = \frac{1}{1+e^{-x}} \tag{14}$$

As one of the earliest categorical assessment models, the logistic regression model is able to determine the result of an operation as 1 or 0 by the given data, and the output variable of this model is always between 0 and 1. The logistic regression model assumes that the dependent variable follows a Bernoulli distribution. Under dichotomous conditions, a set of weights is assumed, and the weighting method is used to calculate, using test samples entered into the logistic regression model, the following.

$$h_0(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_r x_r \tag{15}$$

The hypothesis function of the logistic regression model can be expressed as:

$$h_0(x) = g(\theta^T x) \tag{16}$$

The ROC curve is a visual tool for evaluating classification models to verify the predictive effectiveness of the model and its ability to capture a small number of classes. The ROC curve evaluates the predictive effectiveness of the model through the AUC (Area Under the Curve), where a larger AUC value indicates a better predictive effectiveness of the model. The evaluation criteria are as follows:

The classifier is most effective when AUC=1.When 0.5<AUC<1, the classifier outperforms random prediction, and larger AUC values indicate better prediction. When AUC = 0.5, the classification effect of the classifier is consistent with the random prediction effect, proving that the classifier has no value. When AUC < 0.5, the classifier is worse than random prediction.

In order to investigate the relationship between total sales volume and cost-plus pricing of vegetable categories, we use the multiclassification model in logistic regression to categorize the total sales volume of vegetable categories according to 0, (0,1000), (1000,10000), (10000,30000) in Problem I. The total sales volume of the vegetable categories are categorized according to 0, (0,1000), (1000,10000), (10000,30000) in Problem II. Observing that the cauliflower category and eggplant category contain a small number of vegetable items and their correlation is high, this paper combines them and jointly analyzes the relationship between their total sales and cost-plus pricing. The logistic regression results are shown in Table 4:

Table 4: Classification evaluation table

kind	Foliage	Cauliflower and Eggplant	Edible Mushroom	Pepper	Aquatic Roots
accuracy	0.646	0.667	0.718	0.698	0.778
AUC	0.861	0.793	0.878	0.857	0.875

From the above data, it can be obtained that the accuracy rate of each category is greater than 0.64, and the AUC index is greater than 0.79, indicating that the model is excellent.

3.2 Forecasting the maximum total daily replenishment volume

Firstly, based on the sales volume of each vegetable item from 2020-7-1 to 2023-6-20 to forecast the daily sales volume in the coming week (July 1-7, 2023) in order to find out the market demand in a week, this paper uses the SARIMA seasonal time series to forecast.

Due to space constraints, we use SARIMA for demand forecasting as an example of net coupling (1), and the ACF and PACF of the model residuals are shown in Figure 4:

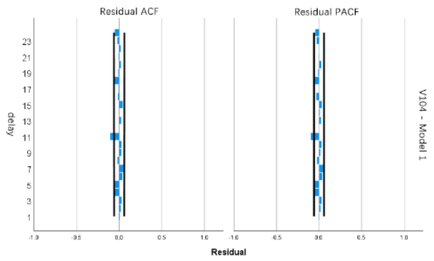


Figure 4: Plot of SARIMA predicted residual ACF, PACF

From the ACF and PACF plots of the residuals in Figure 4, it can be seen that the series are no longer significantly correlated; in addition, the residuals have a Q(18) = 15 with a p-value of 0.003, which indicates that the established SARIMA model is appropriate.

The out-of-sample one-step prediction results of the SARIMA model are shown in Table 5:

Table 5: Table of SARIMA prediction results

Date	Net root (1)
2023.7.1	-0.284648
2023.7.2	-0.430417
2023.7.3	-0.497527
2023.7.4	-0.527875
2023.7.5	-0.544957
2023.7.6	-0.559636
2023.7.7	-0.575762

3.3 Develop pricing and replenishment decisions for vegetables

This problem requires the study of the daily replenishment and the development of pricing strategies for each vegetable category for the coming week (July 1-7, 2023) that makes the superstore the most profitable. In this paper, a number of factors such as discounts, cost-plus pricing, purchase price, sales, and wastage rate are considered and a nonlinear program is established to solve the problem with the objective function P is:

$$\mathrm{P} = \sum_{\mathrm{i}=1}^{251} \mathrm{P_i} \tag{17}$$

$$P_i = (1 - l_i) * y_i * [w * x_i + (1 - w) * x_i * discount] - y_i * x_{i_0} \tag{18}$$

In this paper, the following factors are taken into account when considering the constraints of the objective function: (1) Supermarket capacity constraint: the total daily replenishment quantity of vegetables should be less than the maximum value of the sum of the daily vegetable sales volume in the supermarket; (2) individual category quantity constraint: the daily replenishment quantity of a single category of vegetables should be less than the maximum value of the sum of the vegetable sales volume of that category in the supermarket each day; (3) market demand constraint: the daily replenishment quantity of a single category of vegetables should be greater than or equal to the minimum market demand. Based on the above considerations, we give the following constraints:

$$\text{s.t.} \begin{cases} ll_1 < \sum_{i=1}^{251} y_i < ul_1 \\ ll_2 < \sum_{i=1}^{100} y_i < ul_2 \\ ll_3 < \sum_{i=101}^{105} y_i < ul_3 \\ ll_4 < \sum_{i=106}^{124} y_i < ul_4 \\ ll_5 < \sum_{i=125}^{134} y_i < ul_5 \\ ll_6 < \sum_{i=135}^{179} y_i < ul_6 \\ ll_7 < \sum_{i=180}^{251} y_i < ul_7 \\ y_i \le \hat{y} \end{cases} \tag{19}$$

where the upper boundul = [2483.88, 1265.47, 186.16, 296.79, 118.93, 604.23, 511.14], lower limitll = [64.24, 31.30, 0.63, 0.93, 0.25, 6.07, 3.01].$\hat{y}$ denotes the predicted value of the SARIMA model for replenished items.

Gradient descent algorithm is an iterative optimization algorithm for solving unconstrained optimization problems, especially for dealing with parameter derivable machine learning and deep learning models. In the gradient descent algorithm, an initial point is first chosen, and then the current solution is updated using the gradient of the objective function, causing the solution to move in the negative direction of the gradient, and then the gradient is computed again at the new location, doing the same update, and so on, until some stopping criterion is satisfied. For the constrained convex function optimization problem, the projected gradient descent algorithm obtained by adding the projection idea on the basis of the gradient descent algorithm can update the solution according to the gradient descent in each step, and then project the updated solution back to the feasible domain, which can finally realize the solution of the problem.

Considering that this problem is a convex function optimization problem by constraints, the projected gradient descent algorithm is chosen to solve it in this paper. The specific solution steps are as follows:

Step1 Initialization: choose an initial point x_0 and learning rate.

Step2 Calculate the gradient: calculate the gradient of the objective function at the current pointx_i at the current point, calculate the gradient of the objective function.

Step3 Gradient descent: update the current point according to the direction of the gradient to get a new point x'.

$$x' = x_i - \alpha * \nabla f(x_i) \tag{20}$$

Step4 Projection: projectx' project back to the feasible domain to get the new point$x_i + 1$.

$$x_i + 1 = P(x') \tag{21}$$

Here, the$P(.)$ denotes the projection operation, i.e., finding the point in the feasible domain that is closest to thex' nearest point in the feasible domain.

Step5 Judgment: stop the algorithm if the stopping criterion is satisfied, otherwise return to step 2.

Table 6: Total daily replenishment of each vegetable category (kg)

	Foliage	Cauliflower	Aquatic Roots	Eggplant	Pepper	Edible Mushroom
2023.7.1	12.46	9.73	5.37	9.10	11.91	12.46
2023.7.2	10.21	8.49	6.79	6.01	11.58	10.21
2023.7.3	14.85	5.89	9.46	8.38	10.43	14.85
2023.7.4	12.81	9.43	9.48	6.96	12.91	12.81
2023.7.5	14.26	6.38	9.05	6.42	10.76	14.26
2023.7.6	13.13	9.63	5.89	6.34	10.72	13.13
2023.7.7	10.68	5.94	7.45	7.36	11.62	10.68

Table 7: Representative vegetables in each vegetable category and their pricing

	Foliage	Cauliflower	Aquatic Roots	Eggplant	Pepper	Edible Mushroom
representative vegetable	head of lily magnolia	Purple cabbage(1)	Wild Lotus Root(2)	beefsteak eggplant	Green Bell Pepper(1)	black-skinned macrolepiota mushroom
Vegetable pricing (¥/kg)	40.12	10.73	13.45	11.82	18.33	100.21

In this paper, the initial point is chosen to be 0, the learning rate is 0.1, and the total daily replenishment of each vegetable category is calculated as shown in Table 6:

Of these, the vegetables represented in each vegetable category and their pricing are shown in Table 7:

Ultimately, the maximum value of the superstore's return for the coming week is obtained under this total daily replenishment and pricing strategy is ¥15,214.83.

4. Conclusions

This paper uses cosine similarity analysis, SARIMA, multivariate logistic regression, nonlinear programming and projected gradient descent to first study the distribution pattern and interrelationships of the sales volume of each category and single product of vegetables, and then discusses the total sales volume of each vegetable category with cost plus pricing, and based on the results of the discussion, it gives the pricing and replenishment strategies of the vegetable categories and vegetable single products.

In seasonal time series forecasting using SARIMA, a smoothness test is required. Since there are 251 individual items in this paper that require the use of ARIMA assumptions, the workload is extremely high. Using projected gradient descent for nonlinear programming may converge to a local optimal solution rather than a global optimal solution. In the future we can use machine learning methods like LSTM for prediction to reduce the workload. In addition to this, we can use heuristic algorithms such as Genetic Algorithms, Simulated Annealing Algorithms, etc. to find the global optimal solution.

References

[1] Lisha Mao. 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.

[2] Boqiao Yuan, Baodong Li, Zhidan Qin. Research on the transmission mechanism of short-term fluctuations in vegetable prices--Taking six typical vegetables such as eggplant, cucumber and bean curd as examples [J]. Price Theory and Practice, 2023.

[3] Yajie Lu. Research on dynamic pricing problem of high quality fresh vegetables in supermarkets in China [D]. Beijing Jiaotong University, 2010.

[4] Sihong Gu. A study on dynamic pricing of fresh products in H-retailers considering freshness variation [D]. Donghua University, 2023.

[5] Zhang Lu. Hierarchical clustering method based on distance correlation coefficient [D]. Beijing Jiaotong University, 2019.

[6] Jifu Qian. Construction and application of SVR seasonal time series forecasting model [D]. South China University of Technology, 2010.