Optimization of replenishment and sales strategy of vegetable products based on ARIMA time series model and combination sales

Xiaofeng Zhang*

College of Information Engineering, Tianjin University of Commerce, Tianjin, China, 300133

*Corresponding author: zhangxiaofeng_d@163.com

Abstract. Due to the short shelf life of vegetable items, supermarkets usually restock daily based on the historical sales and demand of each item. In order to develop an optimal strategy for replenishment and sales of vegetable items, this study firstly predicted the total daily replenishment in the coming week using ARIMA time series model, and then this study took the hypermarket revenue as the objective function and considered the correlation between vegetable categories in the sales data, and adopted a combination sales strategy to increase the sales volume. This study establishes the combination sales formula, and uses the third-party library function combined with the genetic algorithm to obtain the optimal coefficient combination to maximize the superstore's revenue, and finally constructs the optimization model of pricing strategy for each category of vegetables, and solves the optimal pricing strategy, and ultimately obtains the daily replenishment of total amount of various categories of vegetables in the coming week.

Keywords: ARIMA time series; Combined sales; Restocking of vegetable items.

1. Introduction

General vegetable goods, the freshness period in the fresh food superstore is shorter, and the longer the sales time the worse the vegetable quality. If most varieties are not sold on the same day, they cannot be sold the next day. Therefore, the historical sales and demand for each commodity is the basis for replenishment of supermarkets. Vegetable varieties, different origins, most of the purchase transaction time in the early hours of the morning, so in the absence of a clear understanding of the specific single product and the purchase price of the case, the merchant of the day of the vegetable category for replenishment decisions, the merchant is generally using the "cost-plus pricing" method of pricing, for the shipping loss of commodities and the poor quality of the commodity will be They will also offer discounts for damaged and deteriorated items. Therefore, it is important to analyze market demand for both replenishment and pricing decisions in order to develop the best strategy for replenishment and sales of vegetable products.

The ARIMA forecasting model used in this study is a common traditional time series forecasting model, which is a time series forecasting method proposed by Box and Jenkins in the 1970s, and can also be called the "Box-Jenkins method" [1]. The model is called ARIMA(p, d, q), where p is the autoregressive AR term, d is the number of difference orders needed to turn the original time series into a smooth time series, and q is the moving average MA term. Genetic algorithm, on the other hand, is a global optimization search algorithm that draws on the biological genetic and evolutionary processes in nature. It adopts mechanisms such as natural selection and genetic manipulation, and is able to continuously accumulate knowledge and experience in the search space through generation-by-generation evolution, and adaptively search for the optimal solution of the objective function[2] .

Vegetables, as fresh produce, have temporal and spatial differences in their production and consumption. Due to the special nature of vegetables, transportation costs are high and seasonally variable, while storage costs are also high, which poses certain risks to the market. Imbalance between supply and demand is a common problem. Due to the characteristics of small rural areas in large cities and restricted natural resource endowment in Tianjin, the vegetable supply structure is complicated.Liu Junting summarized the problems of vegetable industry supply through empirical analysis and normative analysis, and put forward countermeasures and suggestions to optimize the supply-side structure of the vegetable industry [3].Biing-Hwan Lin's study pointed out that the increase of income, the improvement of education, the improvement of diet and health knowledge, and the aging of the population are the most important factors in the supply and demand of vegetables. and health knowledge improvement, and population aging all have an impact on the consumption of agricultural products in the United States. He analyzed and researched the extent to which these influencing factors affect the consumption of vegetable products and based on this, he predicted the trend of U.S. vegetable product consumption in the next two decades [4]. When making replenishment decisions, merchants generally use the "cost-plus pricing" method of pricing, and discounts are applied to shipping damage and poor quality goods. In the study of Western Mining Group Finance Co., Ltd. loan interest rates, Song Jiaquan used the cost-plus pricing method, and its loan interest rates were analyzed and applied in detail [5]. Among various forecasting methods, time series forecasting is one of the oldest and most vital basic forecasting methods in history. Tian Xingjian used ARIMA model theory to process the relevant data in the study of carbon emissions trading market closing, and came up with a realistic ARIMA model of carbon futures product pricing, and finally gave a suitable plan to promote carbon futures products and analyzed its development prospects[6]. Beijing, Shanghai and Shenzhen are the three major cities in China where the development of tourism industry is more prominent. In order to predict the number of tourists in these three cities, Li Yue used the ARIMA model to make predictions, and successfully obtained the prediction results of the number of tourists from 2018 to 2022. The predicted data during this period showed an increasing trend, indicating that the prediction model has good accuracy and prediction ability[7] .

Many scholars have done some research in analyzing sales volume with time series[10], but the results coexist with shortcomings. Ma Jingjing et al. studied the marketing of kiwifruit based on the premise of maximizing benefits and developed a pricing strategy by combining bundled sales, and concluded that the use of a combined bundled sales strategy was beneficial for increasing profits [8].The data required for this study is taken from: https://cumcm.cnki.net/cumcm/studentHome/studentHome.

Download the data of each vegetable product and bundle or recommend according to the correlation between the vegetables in the sales data, create a combination of sales strategies to improve sales and user satisfaction, and combined with genetic algorithms to get the optimal coefficient combinations to maximize the revenue of the superstore, constructed an optimization model of pricing strategy of each category of vegetables, and solved the pricing strategy accordingly, and ultimately got the total amount of replenishment of each category of vegetables in the coming week.

For the relationship between the cost-plus pricing and the total sales volume of each vegetable category, firstly, the missing values and outliers are processed, and then the relationship equation between the total sales volume and the cost-plus pricing of each category is established. The R-square was used as the coefficient of determination to measure the statistical measure of how well the regression model fits the observed data. For the total daily replenishment and pricing strategy for the coming week (July 1-7, 2023), this study uses the ARIMA time series model to calculate the total daily replenishment for the coming week (July 1-7, 2023), and applies the constraints to formulate a pricing strategy to maximize the revenue of the superstore.

2. Supermarket revenue maximization pricing strategy model

2.1. Relationship between cost-plus pricing and total sales for each vegetable category

Firstly, the time, the sales volume of the category and the cost-plus pricing of the category are divided into six different vegetable categories, and then the data is again subjected to data preprocessing including missing value processing and outlier processing. It is known that merchants generally use the "cost-plus pricing" method of pricing, then we can establish the relationship between the total sales volume of each category and the cost-plus pricing formula. Let the total sales volume of each category as $S_j(j = 1 \dots 6)$, cost plus pricing function as $C_j(j = 1 \dots 6)$,the relationship between the two formulas:

$$C_j = y_j(S_j) \tag{1}$$

By finding relevant information, we can know: cost price plus pricing = unit cost × (1 + markup rate); unit cost = (total fixed cost + total variable cost) / sales volume; markup rate = (selling price - purchase price) / purchase price × 100%

Let the wholesale price be $w = \{w_1, w_2 \dots w_i\}$, , the sales volume be $s = \{s_1 \dots s_i\}$, the wastage rate be $l = \{l_1 \dots l_i\}$,the unit price be $u = \{u_1 \dots u_i\}$, and the unit cost of each category be $O_j(j = 1..6)$

$$\begin{cases} C_j = O_j(1 + a_j) \\ O_j = w_i + \dfrac{l_i}{100} \\ a_j = \dfrac{u_i - w_i}{w_i} \end{cases} \tag{2}$$

The cost plus pricing function is obtained as:

$$C_j = \left(w_i + \frac{l_i}{100}\right)\left(1 + \frac{u_i - w_i}{w_i}\right) \tag{3}$$

R^2, known as the coefficient of determination, is a statistic used to measure how well a regression model fits the observed data.R^2 can take values between 0 and 1. $R^2$is calculated by assessing the effect of the fit by comparing the difference between the actual observations and the fitted values of the model. A common calculation is to decompose the sum of squares (SST) into the sum of squares of the regression (SSR) and the sum of squares of the residuals (SSE), and then use the following formula to derive the R^2:

$$R^2 = 1 - (SSE/SST) \tag{4}$$

where SST denotes the sum of total squares, which is the sum of squares of the differences between the dependent variable and its mean; SSR denotes the sum of regression squares, which is the sum of squares of the differences between the predicted values of the dependent variable and its mean; and SSE denotes the sum of squares of the residuals, which is the sum of the squares of the differences between the actual observed values and the predicted values of the dependent variable.

After solving the results are shown in Table 1:

Table.1. function expression

function expression	R^2
C_1 = 4.963 × S1 + 1327.959	0.056 (chili peppers)
C_2 = 6.419 × S2 + 823.400	0.233 (Foliar)
C_3= 9.619 × S3 160.539	0.599 (aquatic roots and tubers)
C_4 = 12.973 × S4 + 725.982	0.351 (edible mushrooms)
C_5 = 15.012 × S5 + 164.923	0.600 (Cauliflower)
C_6 = 11.906 × S6 + 119.185	0.514 (Solanaceae)

2.2. Total daily replenishment

The total daily replenishment of each vegetable category for the coming week (July 1-7, 2023) needs to be requested, and let the predicted daily replenishment of each category for July 1-7, 2023 be K_i(i=1~256). Derived using mathematical knowledge: Daily replenishment = daily sales / (1 - wastage rate / 100)

$$K_i = \frac{S_i}{1 - \frac{l_i}{100}} \tag{5}$$

Then calculate the daily replenishment using excel . Calculate the total daily replenishment for the coming week (July 1-7, 2023) based on the ARIMA time series model.The ARIMA model is a commonly used method for time series analysis and forecasting. It is applicable to data with autocorrelation and seasonality and combines components of autoregression, differencing, and moving average. First, time series data need to be visualized and explored for overall trends, seasonality, and other characteristics. The following analysis is based on data from the chili pepper category. The results are shown in Figure 1:

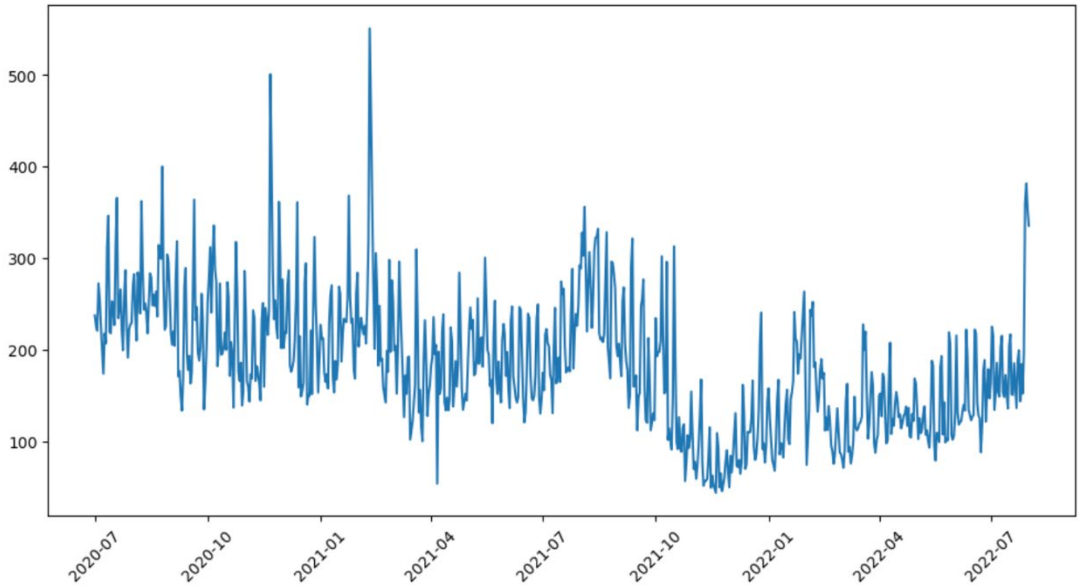

Figure 1. Data map

The data is found to be non-smooth and we need to apply the difference operation to make it smooth. Performing the differencing process, the differencing operation is used to remove the non-smoothness and transform the raw data into a smooth time series. If the data has a trend or seasonality, we can perform first or multiple orders of differencing on the data until we get a smooth series. The order of differencing is denoted by d.

First order differencing can be expressed as: $\Delta Y(t) = Y(t) - Y(t-1)$

The second order difference can be expressed as: $\Delta 2Y(t) = \Delta Y(t) - \Delta Y(t-1)$

The first order, second order difference results are shown in Figure 2:

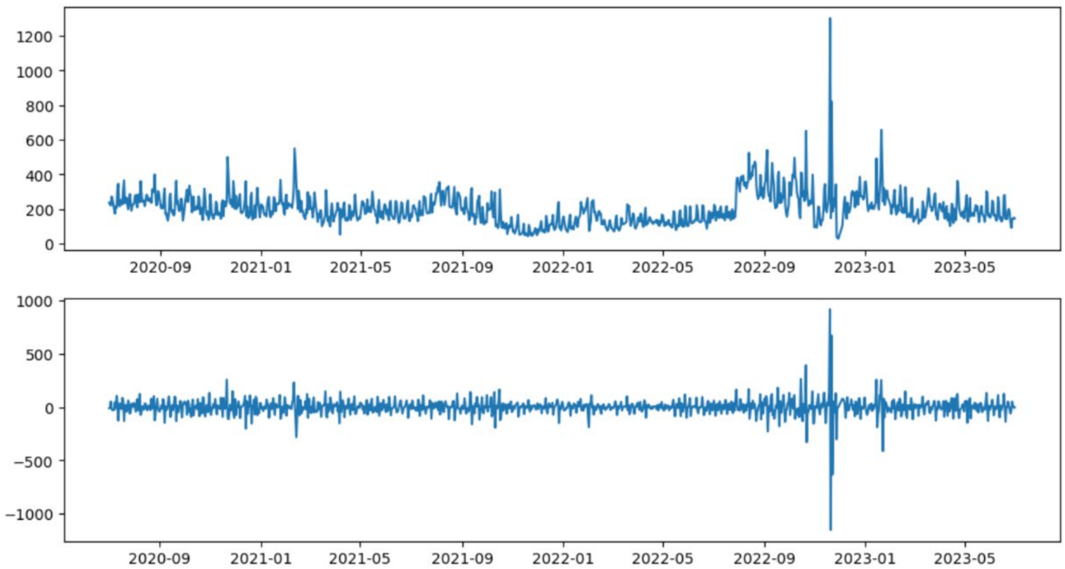

Figure 2. esults of order difference vs. second order difference

Smoothness test: unit root test (e.g. ADF test) is used to determine whether the data is smooth or not [8]. The data was found to be smooth after second order differencing. Then the values of the parameters p and q in the ARIMA model were determined by looking at the ACF and PACF graphs, and based on the determined values of p, d and q, the ARIMA model was fitted. The predicted and true values are shown in Figure 3:

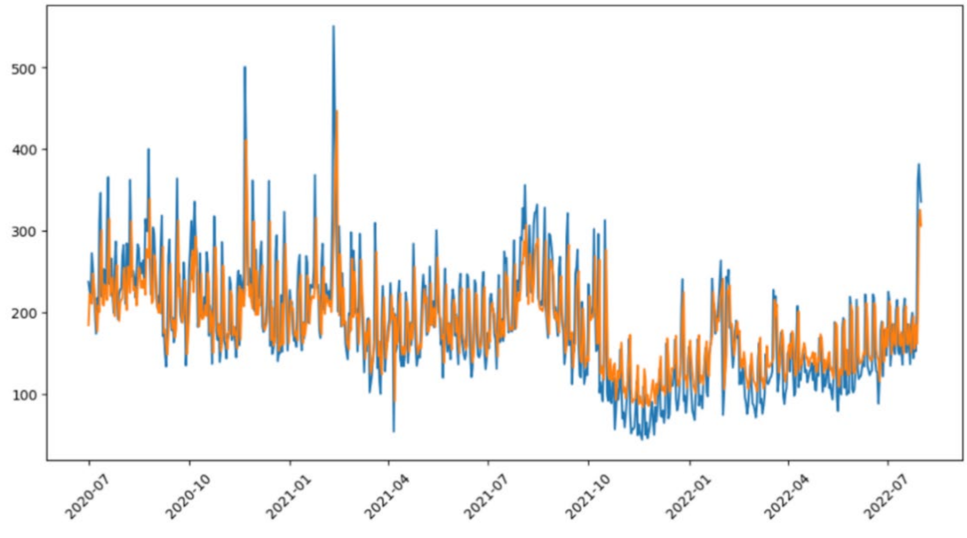

Figure 3. Data fitting plot

After solving the solution the total daily replenishment is shown in Table 2:

Table 2. Total replenishment for the day

Date	Pepper Forecast Results(kg)	Foliar Forecast Results (kg)	Aquatic Roots and Tubers Forecast Results (kg)	Edible Mushrooms Forecast Results (kg)	Cauliflower Forecast Results (kg)	Solanaceae Forecast Results (kg)
7月1日	91.998	141.878	23.286	49.737	25.264	24.119
7月2日	93.555	130.628	23.923	53.219	22.576	22.154
7月3日	94.104	133.325	24.183	54.668	21.302	18.928
7月4日	94.212	136.034	24.437	55.271	20.688	18.517
7月5日	94.218	140.129	24.685	55.523	20.383	17.864
7月6日	94.220	141.704	24.927	55.629	20.223	19.631
7月7日	94.230	141.001	25.164	55.674	20.129	20.659

2.3. Pricing strategy

(1) Preparation of the model

Let the total profit from July 1-7, 2023 be F

$$F=\sum_{J=1}^{\sigma} S_j a_j=\sum_{J=1}^{\sigma} f_j(C_j)a_j=\sum_{J=1}^{\sigma} f_j\left[\left(w_i+\frac{l_i}{100}\right)\left(1+\frac{u_i-w_i}{w_i}\right)\right]\frac{u_i-w_i}{w_i} \tag{6}$$

To ensure maximum profitability, assume that the daily sales volume is equal to the daily replenishment volume. Then total projected profit = six categories projected daily sales x their respective pricing. To wit:

$$F=\sum S'_{j\alpha}C'_{j\alpha} \tag{7}$$

Mathematical language can be used to describe the correlation between sales data and vegetables using graphs and edges in graph theory. We can represent each vegetable as a node in the graph, and the case of two vegetables being purchased at the same time in the sales data is represented as an edge between two nodes.

The specific model is shown in Figure 4:

Input the adjacency matrix A of vegetable types and purchase relationships in the sales data, and a predetermined correlation threshold. Based on the adjacency matrix A, find all edges that satisfy the correlation threshold and construct a correlation graph G. In the correlation graph G, find all maximal complete subgraphs using the maximal clustering algorithm, i.e., find complete subgraphs in which all nodes are connected two by two and no other node can join. The final output, the packages or recommendation combinations in the sales recommendation model. Combine the vegetables in each maximal complete subgraph into a package or make a recommendation.

In order to achieve the goal of maximizing the supermarket's revenue, the following constraints can be considered to achieve "combined sales": from the first question, it can be seen that chili peppers, foliage and mushrooms have the highest correlation coefficients, and they can be sold in bundles: assume that the sales volumes of chili peppers, foliage and mushrooms are S1, S2 and S4, respectively, and then a coefficient k is set to represent the proportion of their bundled sales. coefficient k to denote the proportion of their bundled sales.

The constraints can be defined as:

$$S_1 = k(S_2 + S_4)$$

where k is the proportion of bundled sales. The constraint of sales volume relationship between chili and other vegetables can be set a to represent the sales volume relationship between chili and other vegetables, similarly, the constraint of sales volume relationship between foliage and other vegetables can be set b to represent the sales volume relationship between chili and other vegetables, and the constraint of sales volume relationship between edible mushrooms and other vegetables can be set c to represent the sales volume relationship between chili and other vegetables. These constraints can be adjusted according to the actual situation to ensure that the hypermarket revenue is maximized.

By finding: k=1.5,a=b=c=2

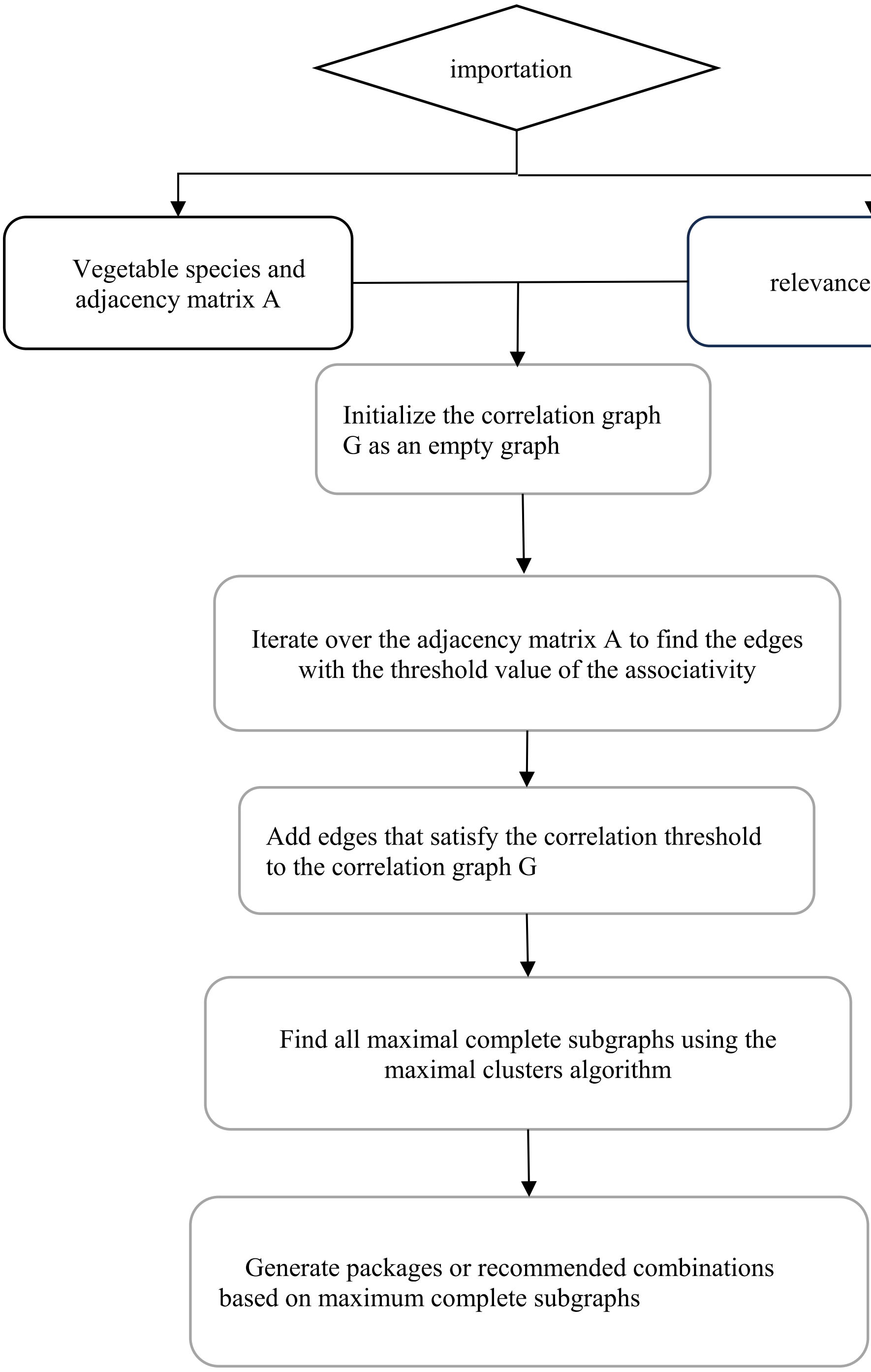

Figure 4. Figure Model building flowchart

(2) Modeling

Objective function:

$$F = \sum S'_{j\alpha} C'_{j\alpha} \tag{8}$$

Upper bound LB,lower bound UB. constraints:

$$\begin{cases} LB'_i \le S_i \le UB'_i (I = 1,2,3,4,5,6) \\ S_1 \le a(S_2 + S_4) \\ S_2 \le b(S_1 + S_4) \\ S_4 \le c(S_1 + S_2) \end{cases} \tag{9}$$

3. Results

For the total daily replenishment and pricing strategy for the coming week (July 1-7, 2023), we calculate the total daily replenishment for the coming week (July 1-7, 2023) using the ARIMA time series model. The constraints are applied to develop a pricing strategy that maximizes the superstore's revenue. The predicted maximum daily sales profit as well as sales volume after one week is shown in Table 3:

Table 3. Day Replenishment Total Sales and Maximum Profit

Date	Maximum Profit ($)	Sales volume (kg)
7月1日	17103.05	S_1 =604.231
7月2日	17100.75	S_2 =604.231
7月3日	17107.41	S_3 =604.231
7月4日*	17109.75	S_4 =604.231
7月5日	17110.19	S_5 =604.231
7月6日	17108.84	S_6 =604.231
7月7日	17103.62	S_7 =604.231

4. Conclusions

A data review was first conducted and relevant mathematical formulas were established based on the principles of the cost-plus pricing method. The historical cost-plus pricing was obtained by operating on the historical data of individual products in the data. Secondly, this study cleaned the data, including dealing with outliers and missing values. The cleaned data were reclassified based on category information and analyzed by linear regression using Python, and the expression of the functional relationship between the sales volume of each category and cost-plus pricing was obtained. Based on the available data, a mathematical formula for the total daily replenishment was established, and the historical daily replenishment volume was obtained. The total daily replenishment for the coming week was predicted using the ARIMA time series model. Then, using the sales volume of each category as the decision variable and the revenue of the superstore as the objective function, and considering the correlation between vegetable categories in the sales data, a combination sales strategy was used to increase the sales volume, and a combination sales formula was established, and the optimal coefficient combination to maximize the revenue of the superstore was obtained by using the third-party library functions in Python and combining with the genetic algorithm.

References

[1] Kenney RT, Frech SA, Muenz LR, et al. Dose sparing with intradermal injection of influenza vaccine[J]. N Engl J Med, 2004, 351(22):2295-2301.

[2] Li MQ, Kou JS. Basic theory and application of genetic algorithms [M]. Beijing:Science Press,2002:32-37.

[3] Liu JT. Research on optimization of supply side structure of vegetable industry in Tianjin[D]. Tianjin Agricultural College,2021.

[4] Biing-Hwan Lin. Fruit and Vegetable Consumption Looking Ahead to 2020[J]. American Journal of Agricultural Economics,2004,65(3):31-36.

[5] Song Jiaquan. Research on the application of cost-plus pricing method based on the loan interest rate of Western Mining Group Finance Company Limited[D]. Lanzhou University,2018.

[6] Tian Xingjian. Carbon futures product design based on ARIMA model[D]. Anhui University of Finance and Economics,2023.

[7] Li Yue. Forecasting the number of tourists based on ARIMA model [D]. Jiangxi University of Finance and Economics,2020.

[8] Peng Weixiang. Problems and method improvement of DF unit root test in time series[J]. Statistics and Decision Making,2022,38(21):53-56.

[9] Jingjing Ma,Tao Yang,Ansheng Ye et al. Research on kiwifruit portfolio sales and pricing strategy based on profit maximization[J].Anhui Agricultural Science,2019,47(08):219-221+227.

[10] Xing Miaoying. Research on flight quality in landing phase based on time series analysis method[D]. Civil Aviation Flight Academy of China,2023.