Research On Supermarket Management Strategies Based on Statistical Inference and Genetic Algorithm Technology

Shaobin Wang *, Heli Geng, Tongshu Jin

Ocean University of China, Qingdao, China, 266100

* Corresponding Author Email: wangshaobin@stu.ouc.edu.cn

Abstract. As one of the indispensable foods in people's daily life, reasonable pricing and replenishment decisions of vegetables are crucial for merchants to achieve the business objectives of lower cost and profit maximisation. Based on this, this paper adopts the commodity information and water flow detail data of a supermarket in Qingdao City from 2020 to 2023 to study the optimal pricing strategy of merchants. The research results show that: through correlation analysis and K-means clustering analysis, there is a certain connection between the six categories of vegetables and 251 single products; through HP filter analysis and visual observation, the sales of vegetable categories and single product sales show a cyclical pattern of change in time; from the seven categories of vegetables obtained from the clustering of each of the seven categories of vegetables, a single product is selected, and through the use of logarithmic nonlinear regression model, it is found that with the increase of the cost of cost plus Pricing increases, the total number of sales decreased logarithmically; finally, through the ARIMA time series prediction model, the sales volume of the next seven days to make a prediction, and then through the genetic algorithm optimisation model, to give the superstore's optimal pricing strategy for the next week. In this paper, through a more scientific and systematic mathematical method to study and analyse the vegetable market, a dynamic and flexible pricing of vegetable commodities as well as replenishment strategies are formulated, which will help the business to continue to revitalize and achieve higher profits.

Keywords: K-means clustering analysis, HP filter analysis, ARIMA time series prediction model, Genetic algorithm optimisation model.

1. Introduction

With the rapid development of social economy and the improvement of people's living standards, China's consumer market shows a trend of continuous active and steady growth. And fresh products as a guarantee of people's livelihood, maintain social stability of the just-needed consumer goods, the market size is huge.

As one of the indispensable foods in People's Daily life, vegetables themselves have the characteristics of high frequency, precision and other consumption, and the market prospect is broad. However, due to the natural properties of vegetables such as periodicality, seasonality, perishable and vulnerable, as well as the different purchase costs caused by a large number of vegetable varieties and different origin, reasonable pricing and replenishment decisions are crucial for merchants to achieve the business objectives of low cost and maximum profit. It is also a matter to consider the promotion plan and sales mix for the demand and supply of vegetables in different time periods. In order to better meet market demand and achieve good economic and social benefits, it is necessary to decide production and sales activities according to the actual situation. It can study and analyze the vegetable market through a scientific and systematic mathematical method, develop dynamic and flexible vegetable commodity pricing and replenishment strategies, and plan reasonable marketing programs according to the analysis results, so as to continuously revitalize business vitality and bring higher profits [1].

2. Data sources

The data in this paper are from the sales records of a supermarket in Qingdao, Shandong Province. The data cover the information of 251 kinds of goods in 6 vegetable categories distributed by the supermarket. Distribution details and wholesale price data of each commodity from July 1, 2020, to June 30, 2023; And the recent rate of depletion data of each commodity.

3. Research method

3.1. Interrelationships and distribution patterns of sales volume of vegetable categories and individual products

(1) Correlation analysis based on sales volume of vegetable categories

We obtained correlation coefficients for the sales volume of the six types of vegetables and constructed a matrix of correlation coefficients:

$$\begin{bmatrix} 1 & \cdots & \mathrm{r_{1n}} \\ \vdots & \ddots & \vdots \\ \mathrm{r_{n1}} & \cdots & 1 \end{bmatrix} = \left(\mathrm{r_{ij}}\right)_{\mathrm{p\times p}} \tag{1}$$

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

$(r_{ij})_{p*p}$is the correlation coefficient of the two variables,x and y represent two different vegetable categories.

(2) Interrelationship analysis between the sales volume of vegetable individual items based on K-means clustering

Aiming at the connection between different vegetable single items, K-means algorithm is used to cluster vegetable single items into clusters. The basic idea of K-means clustering is to cluster according to the prototype, that is, to find out several prototype carvings, or called initial centres, in vegetable single item points. The Euclidean spatial distance to the initial centre is sought for each vegetable single product point, and the sample points are grouped into the class with the smallest distance, and this is iterated for many times, and the mean value method is used to innovate the centroid of each cluster in the course of each iteration until the squared error is reduced to an acceptable value.

(3) Distribution law of vegetable category sales based on HP filtering method

In this paper, HP filtering method is used to analyse the distribution pattern of the total monthly sales of six categories. The core idea of HP filtering method is to use the Hilbert transform to convert the raw data into analytical signals [2-3]. The analytical signal is a complex signal, including the amplitude and phase information of the original signal, which can better describe the periodic characteristics of the signal. Once the analytic signal is obtained, the periodic pattern can be easily analysed.

(4) Analysing the distribution pattern of vegetable single product sales through data visualisation

Through vegetable single product clustering, the single products have been divided into k categories based on the similarity of sales and price, so one single product from each category is randomly selected as a representative, and the sales distribution law of this category is obtained by calculating the average daily sales of this single product in each time period on weekdays and double holidays.

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

(1) Data pre-processing

The formula for cost-plus pricing was obtained by reviewing the literature as follows:

$$Cs = c + \frac{s-c}{c} \cdot c \tag{3}$$

Cs is the cost-plus pricing of a single item, and c is the cost of a single item, i.e., the wholesale price of a single item, and s is the selling price of a single item.

In order to make the results more rigorous, the average of the wholesale price of cost at all moments in the history of a single product is taken as the cost of a single product; similarly, the average of the selling price at all moments in the history of a single product is taken as the selling price of a single product.

(2) Non-linear regression-based analysis of the relationship between each category and cost-plus pricing

For the total sales data that has been processed and cost-plus pricing, divided by category, to analyse the interrelationship between the two. In this paper, non-linear regression is used to fit the data, and in the process, a variety of non-linear models are considered, and the one with the best fitting effect is obtained through comparison.

Any two data sets from the vegetable category were selected and fitted by the following three non-linear models:

$$y = a + b * \ln(x) \tag{4}$$

$$y = \frac{a}{1+e^{b+c*x}} \tag{5}$$

$$y = a - \ln(1 + b * e^{-c*x}) \tag{6}$$

(3) ARIMA time series-based forecasting model for replenishment volume of each category

In order to meet the balance of supply and demand, the superstore will purchase goods in accordance with the daily average of the sales volume of the goods in the past period of time, but there is a gap between the actual daily sales volume in the future and the average value in the past, and this gap is the replenishment volume [4]. The ARIMA (p,d,q) model can transform the non-stationary time series into a stationary time series, therefore, this paper adopts this model to forecast the sales volume of each category in the coming week [5-6].

The ARIMA (p,d,q) model is divided into three parts: in the AR (p) model, p represents the lag order, which indicates that the current sales volume is related to the sales volume of the previous p moments; in the I (d) model, d represents the difference order, which indicates that d times of difference are carried out; and in the MA (q) model, q represents the moving average order, which indicates that the current sales volume is related to the prediction error of the previous q moments.

(4) Genetic Algorithm Optimisation Model for Replenishment Volume of Each Category

In order to maximise the revenue of the superstore, this paper uses a genetic algorithm to solve the problem [7-9]. The basic idea of the algorithm is to model the problem as a population genetic evolution process, in each generation, through selection, crossover and mutation and other genetic operations, to simulate the genetic process in nature, so that the individuals in the population continue to evolve, so as to find the optimal solution to the problem [10-11].

Step1: Objective function construction

$$w = \sum_{i=1}^{6} (x_i - c_i) \cdot y_i - \sum_{i=1}^{6} m_i \cdot c_i \cdot y_i \tag{7}$$

w is the profit, x_i is the cost-plus pricing of the ith category, c_i is the cost of the ith category, m_i is the attrition rate of the i-th category, and y_i is the sales volume of the i-th category.

In this paper, the mean value of the cost-plus pricing of a single product in the vegetable category is taken as the cost-plus pricing of the category; the mean value of the wholesale price of a single product in the category is taken as the cost of the category; and the mean value of the wastage rate of a single product in the category is taken as the wastage rate of the category.

The relationship between sales volume and cost-plus pricing is introduced here, and from the related literature, the relationship between product sales volume and product price can be explained by the price elasticity of demand model:

$$y = i_0 + \ell * \Delta p \tag{8}$$

$$\ell = \frac{\partial y}{\partial x} \cdot \frac{x}{y} \tag{9}$$

i_0 is the baseline sales volume and p is the offset relative to the average price.

Based on the above model, the relationship equation between category sales and cost-plus pricing can be obtained by joining the two and solving the differential equation:

$$x_i = \left(\frac{y_i - o_i}{y_i} \right)^{\frac{\Delta p}{o_i}} \tag{10}$$

Substituting this relationship back into the objective function gives the following simplified result:

$$w = \sum_{i=1}^{6} (\left(\frac{y_i - o_i}{y_i} \right)^{\frac{\Delta p}{o_i}} - c_i) \cdot y_i - \sum_{i=1}^{6} m_i \cdot c_i \cdot y_i \tag{11}$$

Step2: Constraints Construction

1. Single category quantity constraints: for each category, a quantity constraint is required, i.e., the replenishment quantity of each category should be less than the maximum value of the category's one-day sales volume in the historical data:

$$y_i - l_i \le \max \tag{12}$$

l_i represents the average value of the daily sales volume of the ith category in previous years.

2. Superstore capacity constraints: the total amount of replenishment should be less than or equal to the maximum value of the sum of the daily vegetables in the superstore:

$$\sum_{i=1}^{6} (y_i - l_i) \le \max \tag{13}$$

3.Market demand constraint: the replenishment quantity of each category should be greater than the market demand:

$$\min \le y_i - l_i, \forall i = 1,2......6 \tag{14}$$

Under this objective function and constraints, this paper takes the sales volume and cost-plus pricing of each individual product as the decision variables, and conducts planning, i.e., pricing strategy, to find the optimal solution that maximises the benefits.

Step3: Initial population seed generation

The population size is set to 300. The optimisation problem contains 5 optimisation variables and the range of values of the optimisation variables is known. A fixed-length binary code is used to represent the individuals in the population, and the length of the coding string for each variable is L=40.

Step4: Individual fitness evaluation criteria

Here the penalty function method is used to deal with the constraints, from the objective function and constraints of solving the optimal solution of the design variables, it can be seen that its value domain is non-negative, so its original objective function can be converted as follows in order to obtain its fitness evaluation criteria:

$$H(x) = F(x) + a \cdot \max(g(x),0) \tag{15}$$

The smaller the value of, $H(x)$ the better its adaptation.

Step5: Designing genetic factors.

Selection operation uses local competitive selection operator; crossover operation uses single-point crossover operator; variation operation uses basic bitwise variation operator.

Step6: Determine the running parameters of the genetic algorithm.

Step7: Termination conditions

Pre-set the maximum number of evolutionary generations and terminate the evolution when it reaches that number.

4. Results and analyses

4.1. Distribution pattern and interrelationship of sales volume of each category and single product of vegetables

(1) Correlation analysis based on sales volume of vegetable categories

Table 1: Correlation coefficient matrix

category	1	2	3	4	5	6
1	1	0.7469	-0.1891	-0.0345	0.6239	0.546
2	0.7469	1	-0.2025	0.0581	0.4255	0.4287
3	-0.1891	-0.2025	1	0.0541	-0.0532	-0.1892
4	-0.0345	0.0581	0.0541	1	-0.1914	-0.4089
5	0.6238	0.4255	-0.0532	-0.1914	1	0.5755
6	0.5460	0.04287	-0.1892	-0.4089	0.5755	1

As we can see from Table 1, there are certain interrelationships among the six classified vegetables. There are positive correlations between some categories and negative correlations between some categories.

(2) Interrelationship analysis between the sales volume of vegetable individual items based on K-means clustering

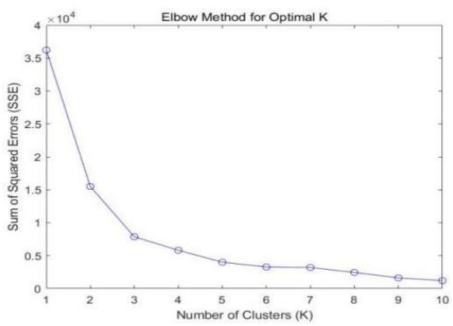

Figure 1: Elbow rule to determine clustering parameter k

From figure 1, Observing the K-SSE graph, when k takes the value of 1-7, the change of SSE is on the large side; when $K \geq 7$,the fold line shows an obvious inflection point, and the SSE turns from a large decrease to a flat decrease, so 7 is identified as the appropriate number of clusters.

The clustering results were visualised using MATLAB:

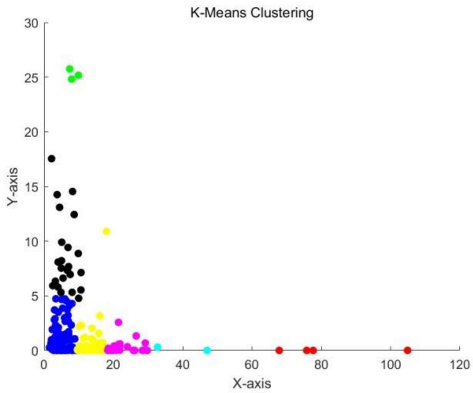

Figure 2: Clustering results for each individual product

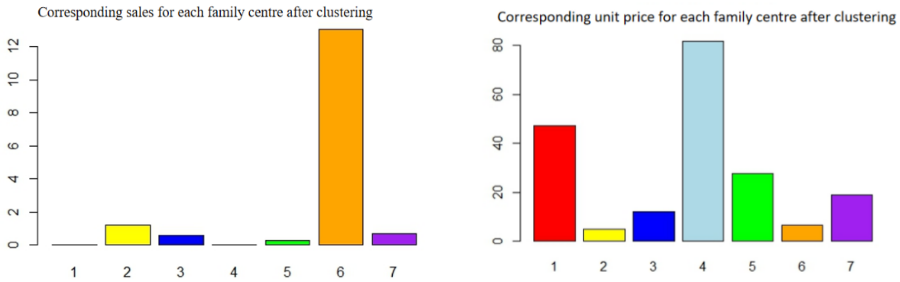

Figure 3: Corresponding sales volume and unit price of each cluster heart after clustering

From the above figure 2 and figure 3, the 251 items can be classified into 7 categories, and the items inside each category have commonality in terms of sales volume and unit price, and the following table 2 illustrates the classification:

Table2: Clustering categories of vegetable individual items

clustering category	Item Name
1	Cow's head lettuce, local small hairy cabbage, etc
2	Yellow cauliflower, chicken fir mushroom, etc
3	Cabbage, pea pods, etc
4	Clean lotus root, broccoli, etc
5	Yunnan lettuce, Chinese cabbage, etc
6	Perilla, Sichuan Red Toon, etc
7	Shanghai green, chrysanthemum, etc

(3) Distribution of sales volume of Vegetables by category in HP Filter Periodic Analysis

The analysis was carried out and the results visualised using EViews. Curve *trend* represents the monthly sales trend over 36 months:

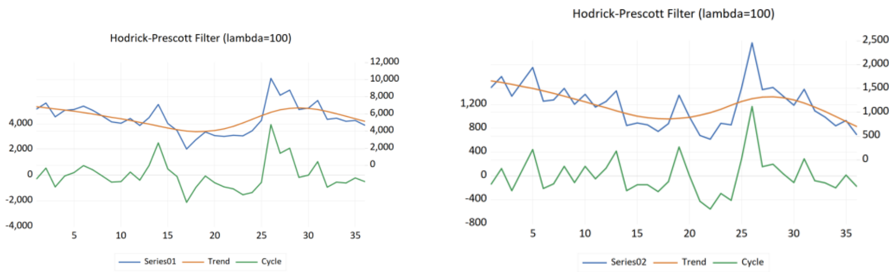

Figure 4: Monthly Sales Trend of Categories 1 and 2

Note: Due to space limitations, only two representative categories are shown here.

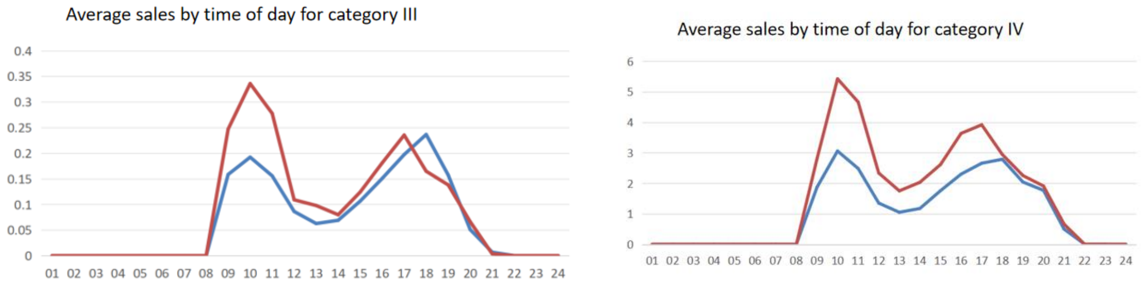

Figure 5: Line graph of daily average sales of individual items in categories 3 and 4 by time period

About figure 5, the red line represents weekends, while the blue line represents weekdays, and the horizontal coordinate is the 24 hours in a day. From the graph, a pattern can be found, subject to the working hours, the vegetable purchasing activity on weekdays is generally lower than that of double holidays. Purchasing activity is less at night, and purchasing activity is mainly concentrated in the daytime.

On weekdays, there are two daily sales peaks for vegetable individual items, respectively 9:00-11:00 and 17:00-19:00. The former peak may be due to the elderly or non-employed people shopping at supermarkets to purchase ingredients for lunch after breakfast; The latter peak may be due to the employed people purchasing dinner ingredients at supermarkets when they return home from work. Looking at the difference between sales on weekdays and holidays, it's easy to see that on weekends, sales of all categories of vegetables are significantly higher than sales during weekday hours.

4.2. Relationship between total sales and cost-plus pricing of each vegetable category and sales strategy

(1) Analysis of the relationship between each category and cost-plus pricing based on non-linear regression

In the three models above, the total sales of a single item is y and the cost-plus pricing is x. By observation, the logarithmic function responds to a certain correlation trend, with total sales decreasing logarithmically as the cost-plus pricing increases.

Therefore, for the other five categories of vegetables, only the logarithmic non-linear regression model was used to present some of the results, as shown below:

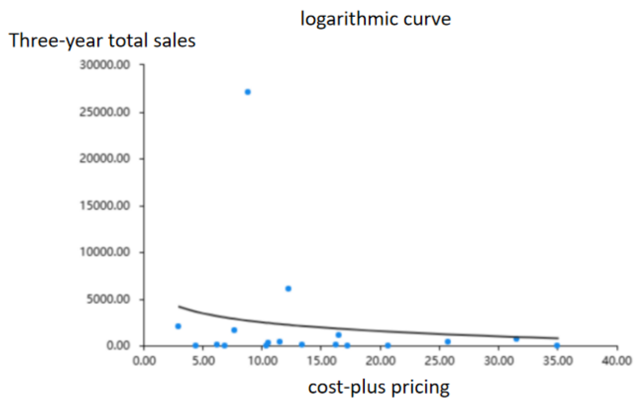

Figure 6: Logarithmic non-linear regression visualisation results

From figure 6, model testing was performed by looking at the sum of squares of the residuals of each model for each of the six categories. If the model's R^2 is close to 1, the regression model is shown to be a good fit. The table below shows the value of R^2 for each category:

Table 3: Regression model R^2 values

Vegetable categories	Model R^2
1	0.09
2	0.184
3	0.02
4	0.018
5	0.044
6	0.032

From table 3, the above R^2 is much smaller than the usual fitted model R^2, indicating that for the non-linear regression model fitted for each category, the fit is not very good and can still be improved.

(2) ARIMA time series based forecast model for replenishment of each category

Use MATLAB to plot the time series, autocorrelation and partial autocorrelation plots of daily sales volume for each category.

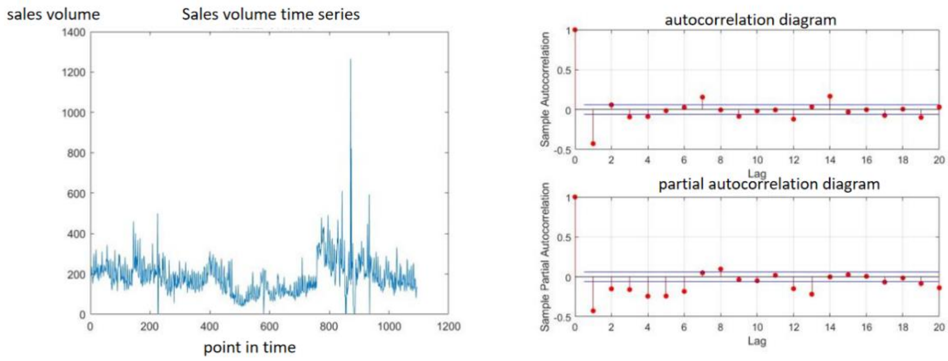

Figure 7: Autocorrelation and partial autocorrelation graphs for category 1 sales volume time series, category 1 sales volume

Note: Because of space reasons, only the correlation graph for the category of the 1st kind is shown here.

Figure 7 from the sales volume time series graph, it was found that the sales volume was unstable and fluctuated in the past, so the 1st order difference was performed; meanwhile, the ACF and PACF graphs were passed with q=1 and p=1. The sales volume was predicted for the next seven days using the fitted ARIMA model (Figure 8).

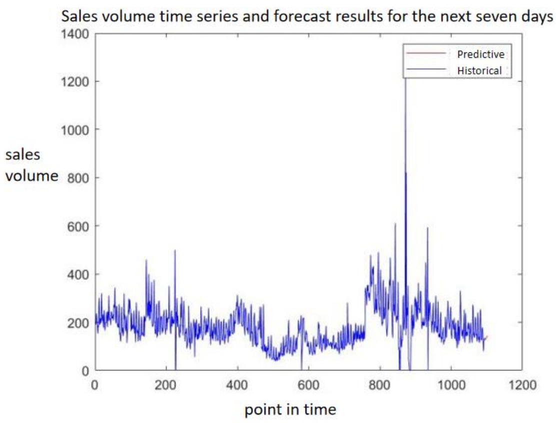

Figure 8: Time series of category 1 sales and predicted sales for the next seven days

To check the goodness of the model, the past sales volume is estimated using the fitted ARIMA model and compared with the processed real data.

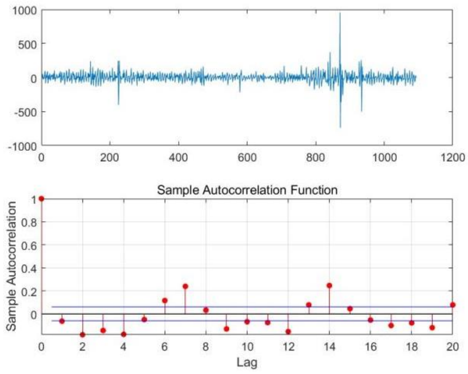

Figure 9: Diagnostic correlation plot of the model

Observing the figure 9, the points fluctuate around 0, indicating a good fit and the model passes the test.

Repeat the above steps for the remaining 5 categories to get the predicted values of total daily sales for the coming week (1-7 July 2023) for each vegetable category (Table 4):

Table 4: Predicted values of total daily sales for each category for the coming week

Vegetable categories / Next day	1	2	3	4	5	6
1	125.83	31.80	15.89	37.54	75.45	30.830
2	131.38	35.53	16.27	34.60	68.63	31.57
3	132.62	39.26	15.44	35.91	61.81	29.61
4	135.7	43.00	14.98	36.09	54.99	28.41
5	138.0	46.74	14.41	36.58	48.18	26.99
6	140.65	50.49	13.86	36.98	41.37	25.62
7	143.16	54.24	13.31	37.41	34.57	24.23

With Δ_{ij} representing the ith category ($i \in \{1,2...6\}$)on day j ($j \in \{1,2...7\}$) with a predicted value of daily replenishment. a_{ij} represents the predicted total daily sales of category i on day j in the coming week. v_i represents the average daily sales volume of category i for the last three years, and the three have the following relationship

$$\Delta_{ij} = a_{ij} - v_i \tag{16}$$

Table5 data will be substituted into the above formula to obtain the predicted daily replenishment for the coming week, the value of 0, that is, the predicted sales volume is less than or equal to the average daily sales in the past, do not need to replenish, the right-most column of the total replenishment of the day.

Table 5: the next week's total daily replenishment of various types of predicted values

Vegetable categories / Next week	1	2	3	4	5	6	Total replenishment amount
1	0.00	0.00	0.00	14.11	0.00	0.00	14.11
2	0.00	1.12	0.00	15.4	0.00	0.00	16.55
3	0.00	4.85	0.00	15.61	0.00	0.00	20.47
4	0.00	8.60	0.00	16.09	0.00	0.00	24.69
5	0.00	12.3	0.00	16.50	0.00	0.00	28.85
6	0.00	16.1	0.00	16.93	0.00	0.00	33.03
7	0.00	0.00	0.00	0.00	0.00	0.00	0.00

(3) Genetic Algorithm Optimisation Model Solution for Restocking Volume of Each Category

The following table 6 lists the values of each parameter for the coming week when making the superstore revenue maximum:

Table 6: Genetic algorithm results

	7.1	7.2	7.3	7.4	7.5	7.6	7.7
y1	1383	1439	1441	1379	1438	1284	1416
y2	215	88.57	165	149.8	189.5	182.3	215.8
y3	185	80.06	164	143.2	148.8	195.5	116.9
y4	109	92.74	95.57	56.78	65.9	96.27	131.9
y5	663	642.1	563.3	413.4	451.8	668.2	462.4
y6	343	385.4	445.1	534.1	532	430.3	400.8
p1	0.03	0.001	0.125	0.007	0.01	0.126	0.062
p2	0.01	0.25	0.032	0.125	0.563	0.063	0.03
p3	1	1	0.542	0.062	0.003	0.5	0.080
p4	0.87	0.003	0.125	0.015	0.073	0.125	0.265
p5	0.15	0.14	0.004	0.25	0.062	0	0.5
p6	0.01	0.062	0.003	0.015	1	0.562	1
profit	21513	20374	21562	20366	21440	21219	30712

The profit of the superstore is maximised when the value of the parameter is taken as taken above. And y_i is the sales volume of each category, to get the cost-plus pricing of each category, it can be solved by the relationship between sales volume and cost-plus pricing.

5. Conclusions

This paper adopts the commodity information and water flow detail data of a supermarket in Qingdao City from 2020 to 2023 to study the optimal pricing strategy of merchants and draws the following conclusions: Firstly, through correlation analysis and Kmeans clustering analysis, there is a certain connection between the six categories of vegetables and 251 single products. Secondly, through HP filter analysis and visualisation observation, the sales of vegetable categories show a cyclical pattern in the monthly dimension, and the single products are generally concentrated in the daytime sales and are divided into two situations: weekdays and double holidays. Third, one single item was selected from each of the seven vegetable categories obtained from clustering, and by using a logarithmic nonlinear regression model, it was found that the total sales volume showed a logarithmic downward trend with the increase of cost-plus pricing. Fourthly, through the ARIMA time series prediction model, this paper makes a prediction of the sales volume in the next seven days, and then optimises the model through genetic algorithm to give the optimal pricing strategy of the superstore for the coming week, which is of great significance for the superstore to better achieve profitability.

References

[1] Vidal, C. J., & Goetschalckx, M. (1997). Strategic production-distribution models: A critical review with emphasis on global supply chain models. European Journal of Operational Research, 98(1), 1-18.

[2] Phillips P C B, Shi Z. Boosting the hodricK-prescott filter[J]. 2019.

[3] Yamada H, Jahra F T. An explicit formula for the smoother weights of the Hodrick–Prescott filter[J]. Studies in Nonlinear Dynamics & Econometrics, 2019, 23(5): 20180035.

[4] Katoch S, Chauhan S S, Kumar V. A review on genetic algorithm: past, present, and future[J]. Multimedia tools and applications, 2021, 80: 8091-8126.

[5] Levenbach, H., & Cleary, J. P. (2006). The Demand Forecasting Handbook. 2nd ed. West Hartford, CT: The Demand Forecasting Company.

[6] Yamak P T, Yujian L, Gadosey P K. A comparison between arima, lstm, and gru for time series forecasting[C]//Proceedings of the 2019 2nd international conference on algorithms, computing and artificial intelligence. 2019: 49-55.

[7] Goldberg, D. E. (1989). Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley Longman Publishing Co., Inc.

[8] Holland, J. H. (1975). Adaptation in Natural and Artificial Systems. University of Michigan Press.

[9] Reeves, C. R., & Beasley, J. E. (2003). Handbook of Metaheuristics. Kluwer Academic Publishers.

[10] Lambora A, Gupta K, Chopra K. Genetic algorithm-A literature review[C]//2019 international conference on machine learning, big data, cloud and parallel computing (COMITCon). IEEE, 2019: 380-384.

[11] Salah A, Çağlar D, Zoubi K .The Impact of Production and Operations Management Practices in Improving Organizational Performance: The Mediating Role of Supply Chain Integration[J].Sustainability,2023,15(20).