ISSN 2616-5805 Vol. 5, Issue 2: 32-40, DOI: 10.25236/AJMS.2024.050206

A mathematical model for automated pricing and replenishment decisions for vegetable items

Chuanyang Zha[1], Guokai Shi[2], Yingying Luo[3]

[1]School of Materials Science and Engineering, Xi'an Shiyou University, Xi'an, 710065, China
[2]Mechanical Engineering College, Xi'an Shiyou University, Xi'an, 710065, China
[3]School of Electronic Engineering, Xi'an Shiyou University, Xi'an, 710065, China

Abstract: *Generally the freshness period of vegetables is shorter, and its sales are greatly affected by time, in order to improve its profitability, so it is particularly important to develop a reasonable replenishment and pricing strategy. In this paper, firstly, the sales data of vegetable commodities were analyzed by descriptive statistics and normality test, and found that the distribution pattern is normal; then Pearson correlation analysis was used to find out the correlation between various categories of vegetables, and it was concluded that the correlation is stronger between cauliflower and foliage, edible mushrooms and aquatic roots and tubers. Vegetable products are the necessities of residents' life, but vegetable products have short freshness period and easy to deteriorate and other problems, so it is necessary to replenish the goods every day, and the goods that have not been bought out should be sold at a discount in a timely manner, in view of the market demand and the interests of the superstore's own needs, so the reasonable replenishment and pricing strategy is also particularly important.*

Keywords: *K-means cluster analysis, Linear regression, LSTM time series modeling, Particle swarm algorithm*

1. Introduction

Cui Ligang [1] et al. constructed a joint replenishment and pricing model for fresh produce based on investment in preservation technology by introducing preservation technology investment parameters based on comprehensive consideration of multi-product replenishment, non-immediate deterioration cycle of stocked goods and pricing, proposed an adaptive differential evolutionary algorithm, and analyzed the parameters in terms of sensibility. Yang Shuai [2] et al. incorporate the quality change of fresh produce into the profit model of retail chain, consider the pricing, replenishment strategy and shelf allocation strategy of fresh food collaboratively, and quickly find the optimal pricing and shelf allocation joint optimal solution by proving with a simple Tiso algorithm.

There are few studies on the replenishment pricing of vegetable commodities, this paper gives the following methods for the pricing and replenishment strategy of vegetable commodities: regression analysis to obtain the linear regression equation between the total sales of each type of vegetables and the cost-plus pricing, the LSTM time series prediction model, which predicts the total sales of each type of vegetables in the week of July 1-7, 2023. A self-constrained optimization model was developed to maximize the benefit revenue for the superstore, and using particle swarm algorithm, the final pricing was solved.

2. Explore vegetable category merchandise sales data

Data Source: National Mathematical Modeling Contest for College Students (mcm.edu.cn) . The data contains information such as vegetable item name, category name, selling time, selling unit price, sales volume, item code, and wholesale price.

2.1 Distribution pattern and correlation analysis of various vegetable categories

2.1.1 Descriptive Statistical Analysis and Normality Test

The total monthly sales of the six major categories of vegetables for the three years were summed and analyzed with descriptive statistics as shown in Figure 1 shows:

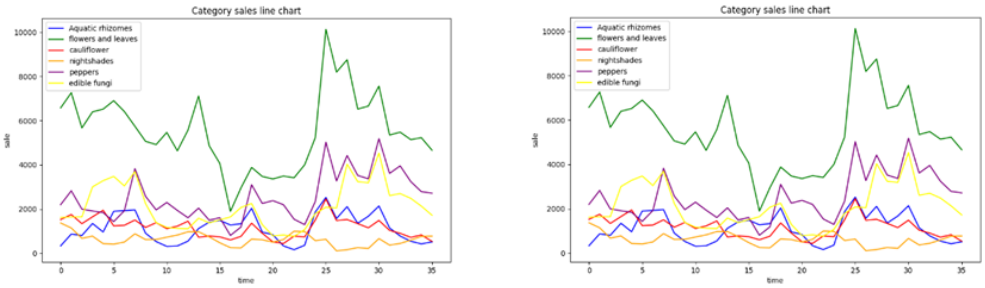


Figure 1: Scatterplot and line graph of three-year sales in six categories

From the figure, it can be seen that there is a big difference between the major categories of vegetables line graph and scatter plot, their sales are all affected by the change of season, in April ~ October each year, sales increased significantly, which may be due to the large number of vegetables marketed in this period of time. In addition, the sales of these six categories of vegetables decreased significantly in November~March each year, which may be due to the low production of vegetables in winter. It is worth mentioning that the sales of these six major categories of vegetables are significantly worse in July 2020~February 2023 than after February 2023, which is due to the blockade of the epidemic.

For the analysis of the distribution pattern among the 6 major categories, it is found that the sales of the flower and leaf category (green data in the figure) are significantly higher than the sales of the other categories of vegetables.

The six categories of vegetables were then analyzed using SPSSPRO and the skewness and kurtosis were as follows Table 1 Shown:

Table 1: Skewness and kurtosis of six major groups of vegetables

variable name	aquatic rootstock	aquatic rootstock	aquatic rootstock	aquatic rootstock	aquatic rootstock	aquatic rootstock
skewness	0.329	-0.970	0.329	-0.970	0.329	-0.970
kurtosis	0.387	0.464	0.387	0.464	0.387	0.464

It can be seen that the median, mean, and standard deviation of the vegetables in the six major categories vary considerably, and further the distribution of sales of vegetables in the six major categories does not overlap over the three-year period.

Since the absolute value of the kurtosis of each type of sample is less than 10 and the absolute value of the skewness is less than 3, the normality test is performed using P-P plots, which respond to whether or not the observed cumulative probability (P) shows a normal distribution by calculating the fit between the observed cumulative probability (P) and the normal cumulative probability (P), and the higher the fit, the more it obeys a normal distribution.

Taking aquatic rhizomes and chemolithotrophs as examples, the results are shown in Figure 2 shown:

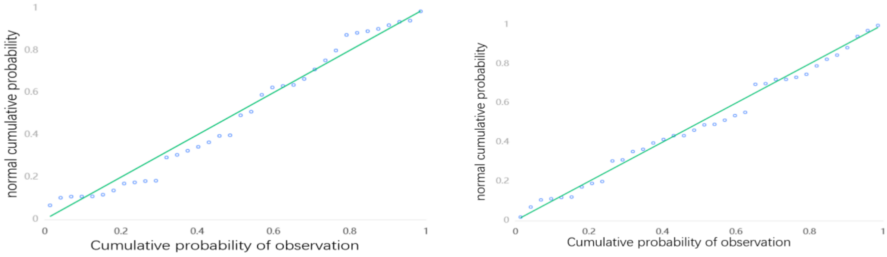


Figure 2: P-P diagram of aquatic rhizomes vs. floral and foliar P-P diagrams

Observation of the normal cumulative probability and observed cumulative probability for each of the aquatic root and leafy vegetable categories reveals that the fit is high, so the sales data for the 2 major categories of vegetables show a normal distribution. Analyzing the sales data of the remaining 4 major categories of vegetables also leads to the conclusion that they show a normal distribution.

2.1.2 Pearson correlation calculation for each category

Pearson's correlation coefficient is applied to continuous quantitative data and the data is required to satisfy a normal distribution. Sales data is quantitative data that satisfies the continuous type and satisfying the normal distribution is also demonstrated in 2.1.1.1.

Normalize the data first:

$$x_{mn}=\frac{x_i-x_{\min}}{x_{\max}-x_{\min}} \tag{1}$$

Pearson correlation coefficient Calculation formula:

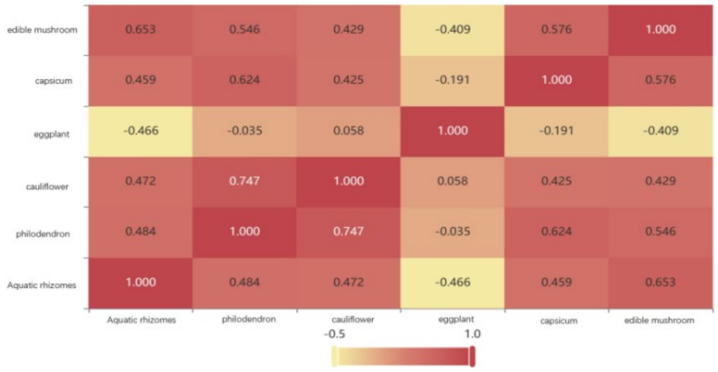
$$\rho_{X,Y}=\frac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y}=\frac{E\left[(X-\mu_X)(Y-\mu_Y)\right]}{\sigma_X\sigma_Y} \tag{2}$$

$$r=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \tag{3}$$

$$r=\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{\sigma_X}\right)\left(\frac{Y_i-\bar{Y}}{\sigma_Y}\right) \tag{4}$$

Where $\frac{X_i-\bar{X}}{\sigma_X}$, $\bar{X}$, σ_X , are X_i sample standardized score, sample mean, and sample standard deviation, respectively.

The heat maps obtained for the six categories of vegetables, as shown in Figure 3:



	Aquatic rhizomes	philodendron	cauliflower	eggplant	capsicum	edible mushroom
edible mushroom	0.653	0.546	0.429	-0.409	0.576	1.000
capsicum	0.459	0.624	0.425	-0.191	1.000	0.576
eggplant	-0.466	-0.035	0.058	1.000	-0.191	-0.409
cauliflower	0.472	0.747	1.000	0.058	0.425	0.429
philodendron	0.484	1.000	0.747	-0.035	0.624	0.546
Aquatic rhizomes	1.000	0.484	0.472	-0.466	0.459	0.653

Figure 3: Heat map of the six categories

The above graph indicates the magnitude of the correlation coefficient values through the color shades, as can be seen from the graph, the eggplant data show negative correlation with the other categories of vegetables, where eggplant has a weak correlation with peppers, cauliflowers, cauliflower, and edibles, which may be caused by the fact that eggplant is usually not mixed with these vegetables during cooking. Stronger correlations were found between cauliflower and foliar, and edible mushrooms and aquatic root vegetables, which had correlation coefficients of 0.6 or more.

2.2 Distribution pattern and correlation of single product sales volume based on K-means clustering

2.2.1 Determine the optimal K value

K-means clustering [3] [4] The analysis requires the magnitude of the K-value, where the elbow rule is used to determine the K-value optimal solution.

By plotting the K-value against the corresponding SSE, it is observed that the curve shows a clear "elbow" bending point. The K-value corresponding to the elbow bend point is considered to be the best choice of K-value. Before this point, increasing the K value significantly decreases the SSE, and after this point, increasing the K value decreases the SSE, which tends to level off, but with a more precise sample delineation.

$$SSE = \sum_{(i=1)}^{k} \sum_{c \in C_i} \left| q - n_i \right|^2 \tag{5}$$

Where q denotes a single sample point and n_i denotes the clustering center to which the sample point belongs.

2.2.2 Analysis of K-valued optimal solutions

The value of K obtained using the elbow rule is shown in Figure 4 shown:

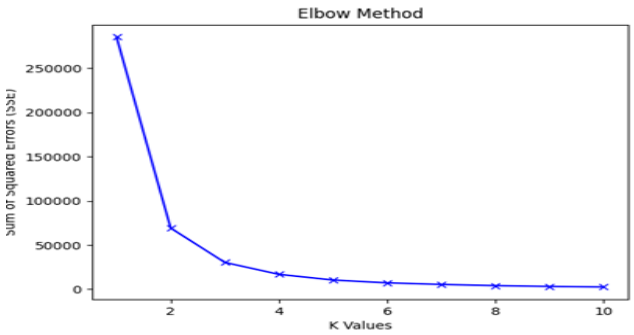


Figure 4: Elbow rule to determine K

As can be seen from Fig. 6, the image slope starts to change abruptly at K=3. As for the K value obtained from the average contour coefficient method, the higher the contour coefficient corresponding to the K value, the better the clustering effect is indicated. Then the optimal value of clustering is K=3, i.e., 254 individual products are divided into 3 categories by considering their sales time and sales volume together.

2.2.3 Distribution pattern of single items after clustering

(1) Descriptive statistical analysis and normality tests

Descriptive tests were performed on the categorized samples using SPSSPRO as shown in Table 2 Shown:

Table 2: Results of descriptive tests for the three categories

variable name	sample size	upper quartile	average value	standard deviation	skewness	kurtosis	S-W test
Category I	866	96.423	242.223	365.108	3.135	15.915	0.66 (0.000***)
Category II	866	68.491	169.179	263.073	3.118	14.114	0.648 (0.000***)
Category III	734	0.013	0.04	0.065	2.809	9.767	0.644 (0.000***)

Where ***, ** and * represent 1%, 5% and 10% significance levels respectively. From the above table, it can be seen that the median, mean, and standard deviation of the three categories are quite different, and their distribution patterns consist of significant differences. Among them, categories I and II have significantly more sales than category III as a whole, but the kurtosis of category III is smaller than that of categories I and II, and the data concentration is high.

(2) Normality tests were performed on the three types of sales data

The significance P-value of the above three categories is 0.000***, a level that presents significance and rejects the original hypothesis, thus the data does not strictly satisfy the normal distribution. However, it can be further analyzed by combining the PP charts with the P-P charts of the three categories obtained, as shown in Figure 5 Shown.

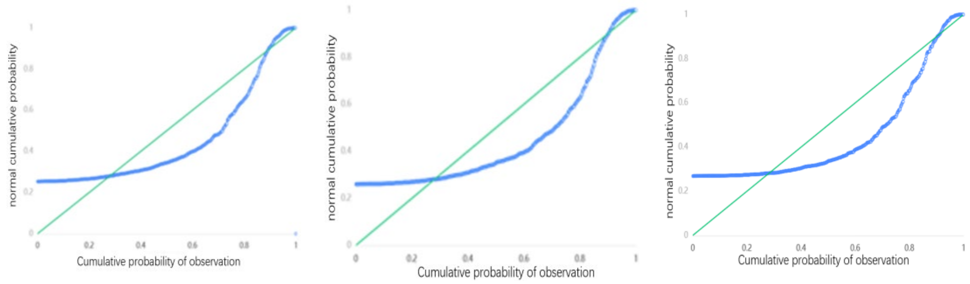


Figure 5: P-P diagrams for the three categories

The figure above illustrates the normality test for the sales data for category 1, category 2, and category 3, showing the fit of the observed cumulative probability (P) to the normal cumulative probability (P), which obeys a normal distribution the higher the fit. Although the fit is not very satisfactory, it is basically acceptable as a normal distribution.

2.2.4 Correlation of the three main classes after clustering

With the same reason as before, to verify that the three types of sales data to meet the continuous type of quantitative data, and require the data to meet the normal distribution, the various types of vegetable sales data to calculate the Pearson correlation coefficient, to get the three major types of data heat map, as shown in Figure 6 .

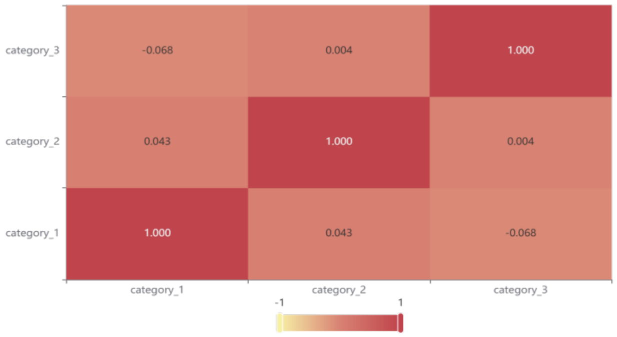


Figure 6: Heat map of the three types of data

It can be seen that category III and category I show a negative correlation and that the correlation between the three categories is weak.

3. A proposal for automated pricing and replenishment strategies for vegetable items

3.1 Linear regression fit

3.1.1 Linear regression modeling

Step1: Calculate the cost-plus pricing. Let the cost-plus pricing of each individual item be X_i , the wholesale price be P_i , the wastage rate be E_i , the selling price be S_i , and the sales volume be D_i , then the cost-plus pricing can be expressed as:

$$X_i = E_i + \frac{S_i - (P_i + E_i \cdot P_i)}{D_i} \tag{6}$$

Step 2: Select the length of time with more data volume, i.e., choose the month as the time unit, summarize the total monthly sales volume of the six vegetable categories, and at the same time calculate the average cost-plus pricing of each vegetable category in each month.

Step 3: Develop a linear regression model of total monthly sales and average cost-plus pricing for each category of vegetables.

$$\begin{cases} Y = X\partial + \zeta \\ v(b) = \dfrac{\sigma}{\sqrt{\sum (x-\overline{x})^2}} \\ v(a) = v(b)\sqrt{\dfrac{\sum x_i^2}{n}} \end{cases} \tag{7}$$

Step 4: Derive linear regression equations for total monthly sales and average cost-plus pricing for each category of vegetables:

$$d_i = ax_i + b \tag{8}$$

Step 5: The data on total monthly sales and average cost-plus pricing for each category of vegetables were fitted and analyzed.

The obtained fit function of total sales of each category of vegetables to cost-plus pricing:

$$\begin{aligned} d_2 &= 8259.786 - 547.43x_2 \\ d_3 &= 2951.434 - 52.065x_3 \\ d_4 &= 943.517 - 47.517x_4 \\ d_5 &= 5730.989 - 394.616x_5 \\ d_6 &= 1910.583 - 92.906x_6 \end{aligned} \tag{9}$$

where d1 to d6 denote the fit function of total sales to cost-plus pricing for cauliflower, foliar, chili, eggplant, edible mushrooms, and aquatic roots and tubers, respectively.

3.1.2 Analysis of linear fitting results - an example of cauliflower species

The regression results of the cauliflower category were analyzed and it was learned that the value of the significance p is less than 0.001, which presents significance, and the VIF is 1, which is less than 10, and there is no problem of multiple covariance, so the model is constructed to satisfy the problem, and the relationship between the total monthly sales of the cauliflower category and the average cost-plus pricing is finally obtained as:

$$d_1 = -116.1x_1 + 2028.005 \tag{10}$$

The relationship between total sales and average cost-plus pricing of cauliflower category shows a negative correlation, i.e., the larger the cost-plus pricing, the smaller its total sales, but due to the small value of R^2, it can be assumed that the negative correlation between the two is less significant.

Similarly, the linear regression equations for the other five categories all demonstrate a less pronounced negative correlation between the relationship between total sales of vegetables and average cost-plus pricing.

3.2 LSTM time series forecasting model

3.2.1 Model building

1) Construct LSTM regression network [5] . Setting the implied unit of LSTM as h, the LSTM model can be expressed as the following equation.

$$\begin{aligned} f_i &= \sigma(W_j \times [h_{i-1}, x_i] + g_j) \\ i_i &= \sigma(W_i \times [h_{i-1}, x_i] + g_i) \\ \tilde{c}_i &= \tanh(W_c \times [h_{i-1}, x_i] + g_c) \\ c_i &= f_i \odot c_{i-1} + i_i \odot \tilde{c}_i \\ o_i &= \sigma(W_o \times [h_{i-1}, x_i] + g_o) \\ h_i &= o_i \odot \tanh(c_i) \end{aligned} \tag{11}$$

2) Back-normalize the prediction result. Let y be the normalized prediction result, and the inverse normalized prediction result, then there are.

$$y_{ar} = y \cdot (x_{\max} - x_{\min}) + x_{\min} \tag{12}$$

3.2.2 LSTM model solving and analysis

By importing the data table of the total sales volume of the six categories of vegetables per day during the period from July 1, 2020 to June 30, 2023 into Matlab and setting the data set of the training set: the data set of the test set = 9:1, the total sales volume of the six categories of vegetables in the coming week can be predicted, and the accuracy of the prediction can also be verified by the prediction of the training set on the test set as well as by the magnitude of the root-mean-square error.

The results were analyzed using aquatic root vegetables as an example, and training set trend plots, sales volume prediction plots for each category of vegetables for the next 7 days, as well as mean square error plots, and training set prediction test set plots for aquatic root vegetables were obtained, as shown in Figure 7 and Figure 8.

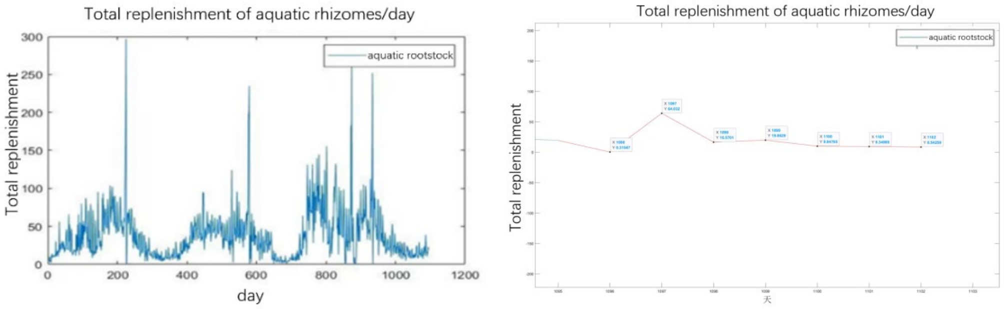


Figure 7: Aquatic Roots and Tubers Training and Trend Chart and Sales Forecast for the Next 7 Days

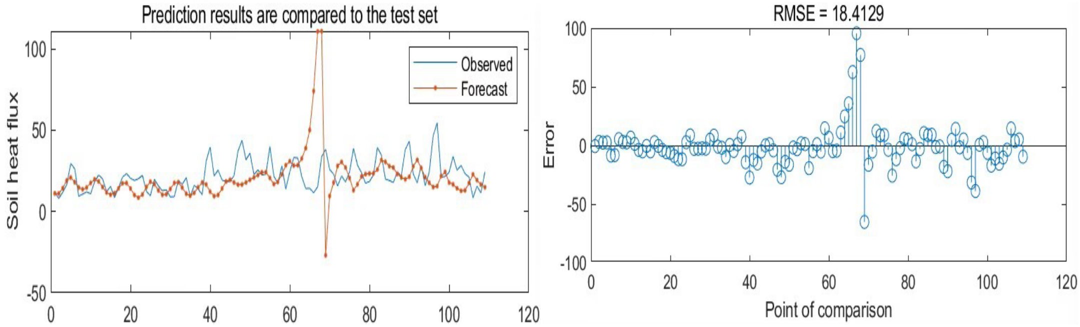


Figure 8: Mean square error plots for aquatic rhizomes and training set prediction test set plots

The analysis of the prediction results of aquatic root vegetables shows that the predicted sales volume (i.e., superstore replenishment) of aquatic root vegetables for each day in the coming week is 0.31, 64.03, 16.57, 19.86, 9.95, 9.35, and 8.54, respectively, and from the running results, the mean-square error value of the aquatic root vegetables, RMSE = 18.412, is also relatively small, so that the model has a good effect on the sales of aquatic root vegetables sales prediction effect is better. Analyzing the prediction results, the predicted value of sales volume on the 2nd day is the highest, probably people have already consumed the vegetables stocked last week and carried out the stocking plan, and the sales volume is less on the last days of the week, which is roughly similar to the actual situation.

The graph of the running results of eggplant, aquatic roots and tubers, cauliflower, cauliflower and leaves, chili peppers, and edible mushrooms and the replenishment for the coming week are as follows Table 3 Shown.

Table 3: Total daily replenishment (kg) for each vegetable category from July 1-7, 2023

dates	eggplant	aquatic rootstock	cauliflower	philodendron	capsicum	edible fungi
2023-7-1	26.79	0.31	32.12	187.68	113.56	69.56
2023-7-2	26.55	64.03	35.28	178.45	105.60	63.58
2023-7-3	20.54	16.57	22.85	125.60	76.35	43.42
2023-7-4	19.86	19.86	25.48	121.15	72.86	37.42
2023-7-5	18.56	9.95	26.53	120.56	73.18	42.75
2023-7-6	19.58	9.35	28.59	116.59	79.65	43.56
2023-7-7	22.11	8.54	33.45	134.25	95.60	54.32

3.3 Self-contained optimization model

3.3.1 Moel building

Based on the information in the question, we model the problem as an optimization problem, i.e., to maximize the benefit revenue for the superstore. We use a self-constrained linear programming model to solve the problem:

$$\max : We = \sum_{i=1}^{6}\sum_{j=1}^{7} \overline{x}_{ij} \cdot (Se_{ij} - 1) - \overline{p}_{ij}(1 + E_{ij}) \cdot Se_{ij}$$

$$s.t.\begin{cases} d_1 = 8259.786 - 547.43x_2 \\ d_2 = 8259.786 - 547.43x_2 \\ d_3 = 2951.434 - 52.065x_3 \\ d_4 = 943.517 - 47.517x_4 \\ d_5 = 5730.989 - 394.616x_5 \\ d_6 = 1910.583 - 92.906x_6 \\ Se_{ij} \le Ce_{ij} \end{cases} \tag{13}$$

where d_1 ~d_6 are the fitting functions between total sales and cost-plus pricing for each major category, respectively.

3.3.2 Model solution

The predicted data is used as the initial value and combined with the particle swarm algorithm to solve the problem, the results are as follows Table 4 shown.

Table 4: Pricing strategies for each vegetable category

all kinds of vegetables	eggplant	aquatic rootstock	cauliflower	philodendron	capsicum	edible fungi
Vegetable pricing (yuan/kg)	6.73	15.55	9.85	6.53	10.23	8.64

4. Conclusions

With the continuous expansion of the vegetable fresh market scale, the competition in the vegetable retail industry becomes more and more intense. In order to help superstores improve their business model, this paper, in response to the market demand as well as the superstores' own interest needs, establishes an LSTM time series prediction model to predict the total sales volume of each type of vegetables in a week, and establishes a self-constrained optimization model to solve the optimal pricing strategy for each type of vegetables using the particle swarm algorithm. Due to the limited vegetable sales data, the training samples are not very sufficient and comprehensive, so there is still a certain degree of error between the predicted value and the measured value, and there is still a certain amount of upside. With the increasing dimensions of the data samples and after rigorous training, the accuracy of the model prediction will become higher and higher.

References

[1] Cui Ligang, Li Yali, Liu Jinxing et al. Joint decision making for multi-product replenishment and pricing considering investment in preservation technology [J]. Industrial Engineering and Management, 2023, 28(03):17-26.

[2] Yang Shuai, Huang Xiangmeng, Wang Junbin. Research on joint optimization strategy of shelf allocation and pricing for fresh food [J]. Supply Chain Management, 2022, 3(08):49-59.

[3] Chen Ming, Li Junxiang, Qu Deqiang et al. Planning of urban emergency logistics facilities based on K-means clustering algorithm - an example of vegetable delivery data at a certain stage in Changchun City[J]. Logistics Science and Technology, 2023, 46(17):57-60.

[4] Liu Quanhong, Tang Fuxing. Optimization of site selection for faulty shared bicycle recycling center based on K-means clustering algorithm and center of gravity method [J]. Operations Research and Management, 2023, 32(07):85-91.

[5] Kowsar T, Zeinab T, Negin D. Ensemble models based on CNN and LSTM for dropout prediction in MOOC [J]. Expert Systems with Applications, 2024, 235.