A Study on Vegetable Commodity Selling Price and Sales Forecasting Based on ARIMA-Differential Evolutionary Algorithm

Chengcheng Liu *, #, Guangyin Zhou #, Caiying Zhong #

School of Biomedical Engineering and Informatics, Nanjing Medical University, Nanjing, China, 211166

* Corresponding author: 13023590348@163.com

#These authors contributed equally.

Abstract. Fresh food superstores need to develop accurate replenishment plans to cope with the pricing challenges caused by the short freshness period, many varieties, and large differences in origins of vegetables, to meet customer demand and to protect the sales program. In this paper, we firstly draw bar charts to show the distribution of sales among the top 10 best-selling and least-selling vegetable items in each vegetable category, and draw correlation coefficients heat maps to show the correlation of sales among the six vegetable categories. Since there are too many vegetable categories, only the 10 best-selling individual items are selected to draw the heat map of correlation coefficient, and hierarchical clustering is performed to study the correlation. Then a cost-plus pricing model is established to represent the recommended selling price through the total cost per unit and profit margin, while the total sales volume of each category is obtained by accumulating the sales volume of each individual item. In this paper, K-means clustering was used to aggregate the six vegetable categories into three broad categories, and it was found that the sales volume and the recommended selling price were negatively correlated. In order to predict the daily replenishment volume of each category from July 1-7, 2023, this paper screens the daily sales volume of the categories in June and July in the past two years, and predicts it with the ARIMA model in time series analysis, and draws line graphs to show the prediction results, and gives the daily replenishment volume strategy at the same time. Finally, differential evolutionary algorithm is used to introduce the optimal selling price, which is 39 yuan/kg for aquatic roots and tubers, 24 yuan/kg for flowers and leaves, 17 yuan/kg for cauliflower, 3 yuan/kg for eggplant, 26 yuan/kg for chili peppers, and 37 yuan/kg for edible mushrooms. The maximum gain was 73,525.45 yuan.

Keywords: Vegetable Commodity, Selling Price, Sales Forecast, ARIMA, Differential Evolutionary Algorithms.

1. Introduction

In fresh produce superstores, the freshness period of vegetable items is short, and the quality will deteriorate with the extension of sales time, and superstores generally restock based on historical sales and demand [1]. Because of the many varieties of vegetables, different origins, and the transaction time is too early, merchants have to make the restocking decision of the day's vegetable category without exactly knowing the specific single product and the purchase price [2]. Vegetable pricing is generally based on the "cost-plus pricing" method [3], and damaged and deteriorated goods are sold at a discount, so market demand analysis is important for replenishment and pricing decisions. Meanwhile, there is a correlation between the sales volume of vegetables and time, and the supply varieties are more abundant from April to October, so the sales mix becomes important.

Lv Wang [4] proposed a fresh vegetable sales trend prediction model based on fuzzy information granulation and improved SVM, combining the fuzzy information granulation method and optimized particle swarm algorithm to improve the support vector machine model, and trend prediction of the sales data of an e-commerce platform in Anhui Province for January 1, 2018-December 31, 2019; Yanliang Zhang et al [5] proposed a fractional-order gray prediction model that predicts the demand for fresh products and introduces Ganfuyuan fresh food e-commerce sales data for validation.

In this paper, the ARIMA model is used to predict the daily sales of vegetables and give the daily replenishment quantity strategy. Then the optimal selling price of each category of vegetables is introduced by using adopting differential evolutionary algorithm.

2. Explore vegetable commodity sales data

Data from: http://www.mcm.edu.cn/

2.1. Preprocessing of data

Due to the large amount of raw data, a missing value test is performed on the data first to confirm that there are no missing values, and then by drawing a box-and-line diagram of the distribution of the total sales volume of the six categories for three years [6] the existence of outliers is found, and the box-and-line diagram is shown in Figure 1.

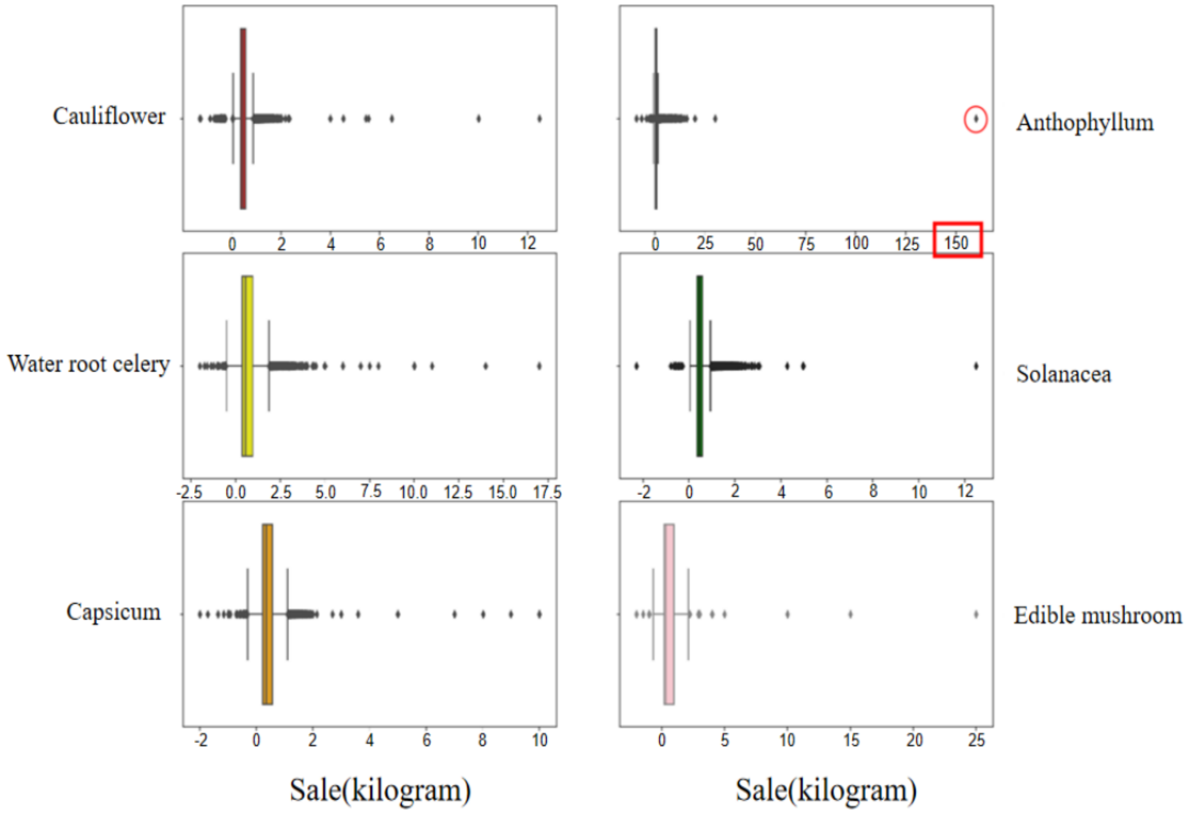

Figure 1. Box line chart of sales by vegetable category

Taking the sales volume of leafy and flowering vegetables as an example, there was only one case of sales volume greater than 150 kg in three years, so it is regarded as an outlier, and this paper stipulates that the upper limit of the normal value of sales volume is 1.32 kg and the lower limit is -0.388 kg according to the statistical method of quartile spacing. So, this paper replaces the outliers that are not in the above range with the median instead of directly deleting them to avoid affecting the subsequent overall probability distribution. For the sales volume where the sales type is returned, it is regarded as negative sales, and this part of the data is retained to ensure the accuracy in the real vegetable sales and purchases, which can be subsequently accumulated with the normal sales volume data.

Then this paper draws a histogram of the distribution of total sales volume of vegetables by category and single product for three years through the updated data, as shown in Figure 2.

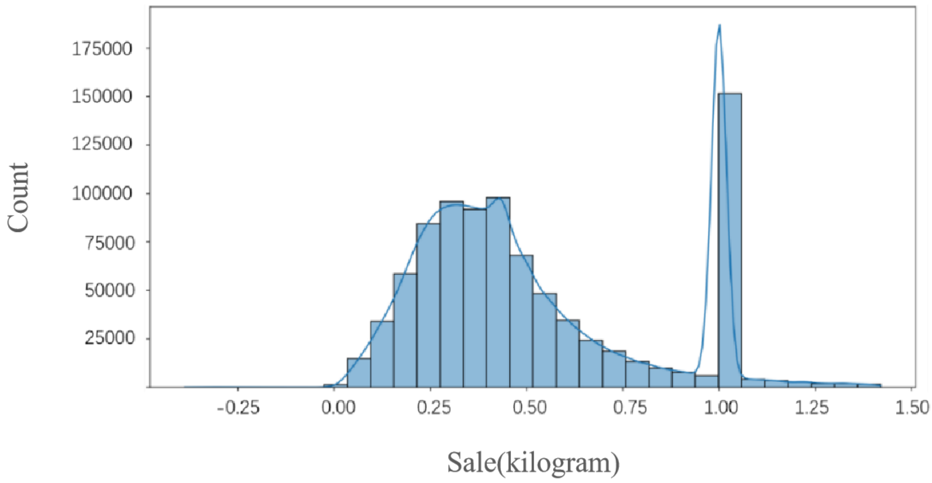

Figure 2. Histogram of sales volume distribution

It was found that some areas of the image conformed to normal distribution, so the skewness was further calculated and the Shapiro-Wilk normality test was performed, and it was rejected that the data conformed to normal distribution with a set significance level of 0.05. Finally, the data in Annexes 1 and 2 were selectively merged based on the single-item code, and the three columns of data, namely, classification name, single-item name and cumulative sales volume, were retained to facilitate the study of the subsequent distribution patterns and interrelationships.

2.2. Distribution pattern of vegetable sales by category and individual item

Using the above processed data, a histogram of the distribution pattern of the sales volume of each category of vegetables was drawn using visualization methods, as shown in Figure 3.

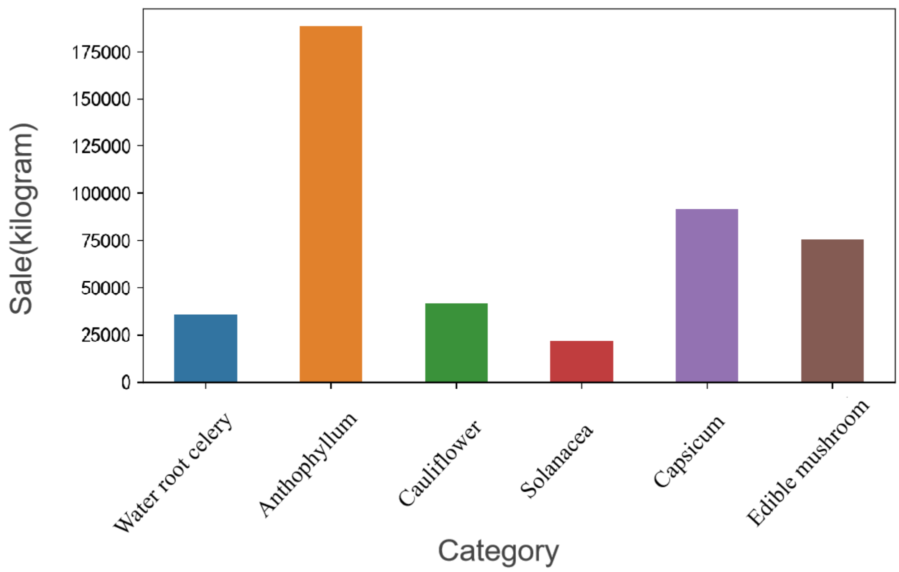

Figure 3. Histogram of Category Sales Volume Distribution

It was observed that the foliage category had the highest sales volume of more than 175,000 kilograms, which was due to the fact that there were more individual species belonging to the foliage category with lower unit prices, and that they were more likely to be purchased for consumption as they were rich in fiber and nutrients. Sales of chili peppers and edible mushrooms are less different, numerically second only to the foliage category, because chili peppers and mushrooms are often used in condiments and soups, and are easy to store for a long time at room temperature. Sales of aquatic roots, cauliflower and eggplant were lower, as customers relatively preferred to buy leafy vegetables that are affordable and rich in nutritional value. Of course, this may also be related to seasonal factors.

For the study of the distribution of sales volume of individual products, due to the large number of types, there are 251 types. Therefore, this paper selects the top 10 best-selling single products and the top 10 least-selling single products for comparative analysis, and draws the histogram of the distribution pattern of the total sales volume in three years respectively, as shown in Figures 4 and 5.

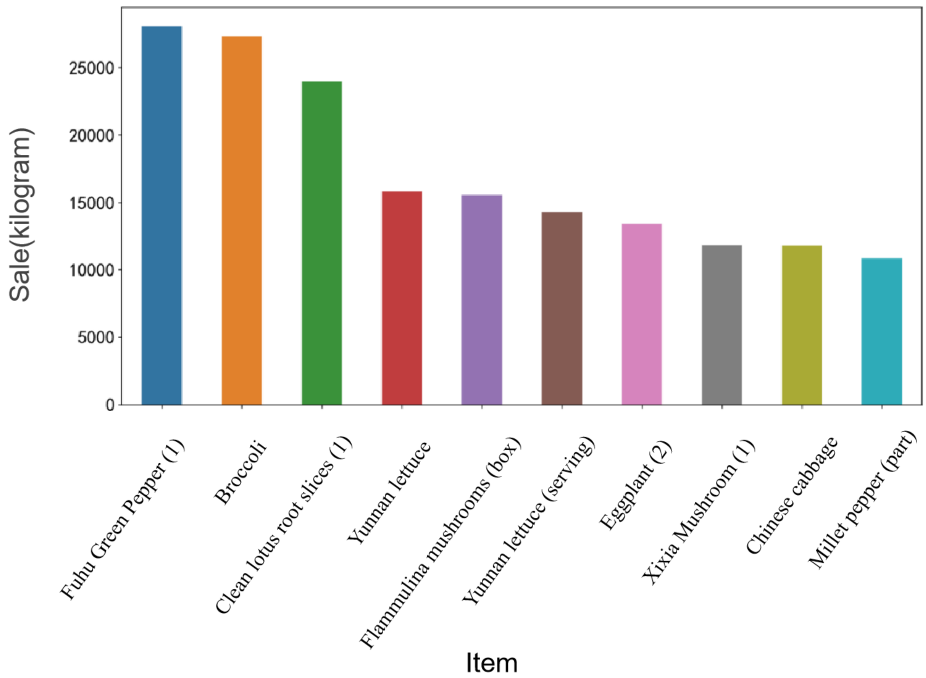

Figure 4. Histogram of sales volume of the top 10 best-selling vegetable individual items

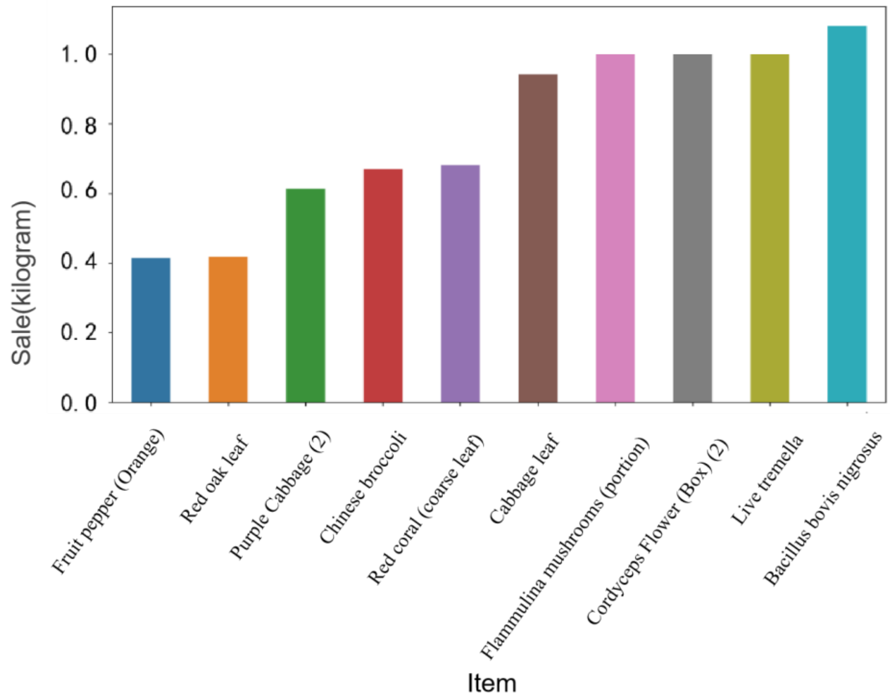

Figure 5. Histogram of individual sales of the 10 least popular vegetables

Observing the two bar charts, it is found that the single items belonging to the flower and leaf category and the pepper category are the best sellers, such as Wuhu green peppers and broccoli, which are in line with people's daily dietary needs, whereas the single items, such as red coral and kale, are sold in small quantities due to less exposure to the people and higher prices.

2.3. Study the correlation between the sales volume of each category and individual product of vegetables

In this paper, a heat map reflecting the interrelationships is drawn through the size of the sales volume of each category. Since the previous test data was found not to conform to normal distribution, the default Pearson correlation coefficient was modified to Spearman correlation coefficient in this paper. As shown in Figure 6.

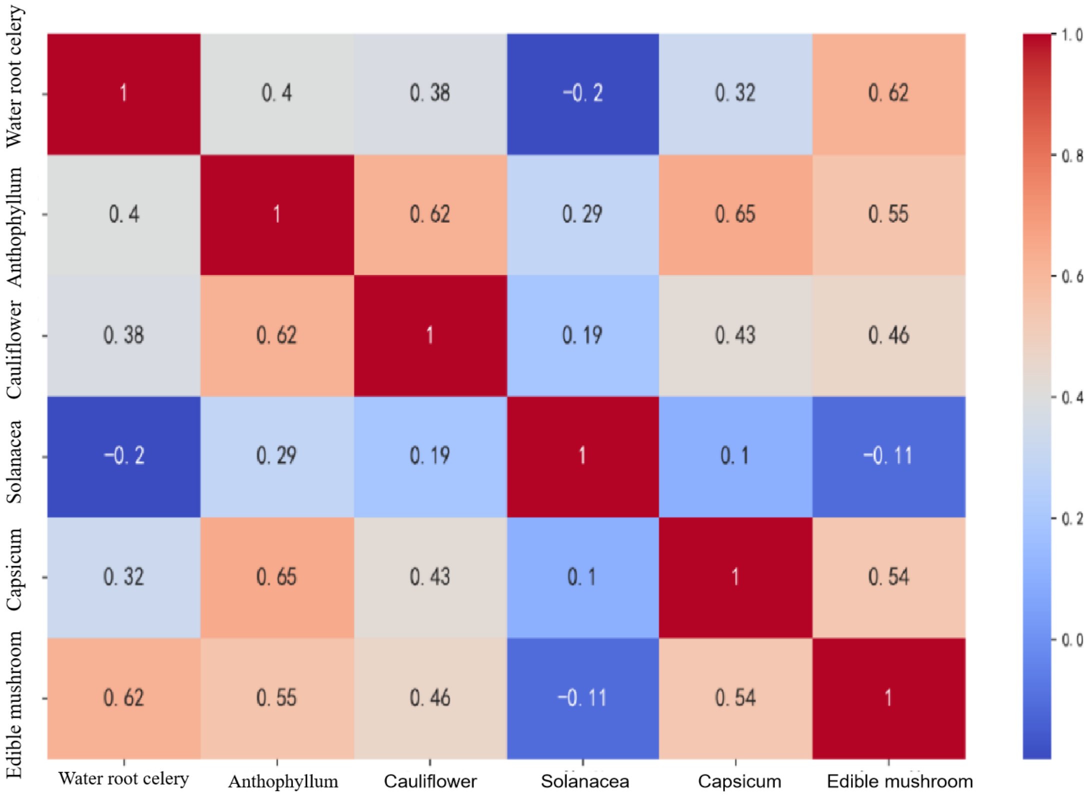

Figure 6. Heat map of sales volume between vegetable categories

Observing the correlation between the sales volume of the six categories, it was found that the correlation between the leafy flowers and the chili peppers was higher, indicating that consumers tend to buy leafy flowers along with chili peppers as an accompaniment for cooking. On the other hand, negative correlations were found between eggplant and edible mushrooms as well as aquatic roots and tubers, which may be related to the inconsistency of production seasons.

For the correlation between the sales volume of individual items, the top 10 best-selling items were selected for this paper because there were too many categories. A heat map reflecting the correlation between the sales volume of the top 10 best-selling individual products is drawn, and hierarchical clustering is carried out on this basis. Hierarchical clustering is a very good technique to find the internal structure of the dataset to understand which individual items are similar in terms of sales and group them together. As shown in Figure 7.

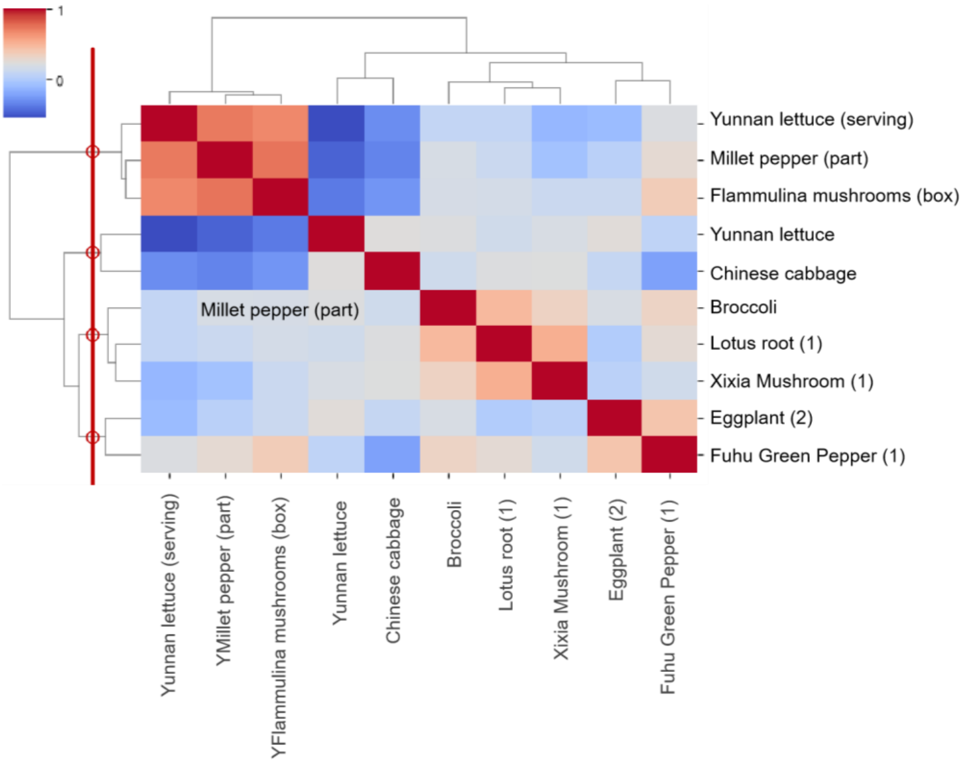

Figure 7. Vegetable Single Product Correlation Chart

It was observed that when the 10 individual items were clustered into 4 categories, the first category was a combination of Yunnan lettuce (portion), millet pepper (portion), and enoki mushrooms (portion), the second category was a combination of Yunnan lettuce and Chinese cabbage, the third category was a combination of broccoli, net lotus root (1), and Xixia shiitake mushrooms (1), and the fourth category was a combination of purple eggplants (2) and Wuhu green peppers (1). This indicates that customers like to add flavorful condiments, such as millet peppers and green peppers, while purchasing vegetables for home consumption.

3. Vegetable commodity selling prices and sales forecasts

3.1. Study the relationship between total sales and cost-plus pricing by category

(1) Cost-plus pricing model

Cost-plus pricing [7] is a pricing strategy that sets prices based on the cost of production and desired profit. The core idea of this strategy is that the price should cover at least the cost of production and distribution and allow for some profit. In this pricing strategy, due to the perishable nature of vegetable products, this paper needs to consider the wastage rate.

Firstly, the wastage rate is converted to decimal form and the formula is calculated as follows:

$$L = \frac{l}{100} \tag{1}$$

This where L is the attrition rate in decimal form, followed by the calculation of the total cost per unit equation below:

$$c = w \times (1 + L) \tag{2}$$

This where w is the wholesale price and then the profit margin r is calculated with the following formula:

$$r = \frac{O - c}{c} \tag{3}$$

The final formula for obtaining cost-plus pricing is as follows:

$$P = c \times (1 + r) \tag{4}$$

This is where P denotes cost-plus pricing, i.e., the recommended selling price. c denotes the total cost per unit, r denotes the markup rate, and O denotes the selling unit price. The results of some of the operations are shown in Table 1 below:

Table 1. Recommended selling prices of individual products

Product code	Item name	Unit price	Attrition rate	Retall prices	Total unit cost	Profitability	Suggested price
102900005117056	hot pepper	7.60	0.07	4.32	4.63	0.64	7.60
102900005115960	bok choy	3.20	0.22	2.10	2.57	0.25	3.20
102900005115823	cabbage	10.00	0.14	7.03	8.04	0.24	10.00
102900005115908	choy sum	8.00	0.14	4.60	5.23	0.53	8.00
...	...	...	...	...	...	...	...
102900005115250	Mushroom	24.00	0.11	15.60	17.28	0.39	24.00
102900011022764	eggplant	12.00	0.07	7.00	7.48	0.60	12.00
102900011016701	paprika	5.20	0.06	3.63	3.84	0.36	5.20

After cost-plus pricing is determined, this paper next calculates the total sales volume of each category by first accumulating the sales volume of individual items, and then finds the average recommended selling price of each category by calculating the mean of the recommended selling price of all individual items under the six categories. Next, this paper first performs K-mean clustering for the 6 categories and finds that some categories can be combined, thus simplifying the decision-making process.

(2) K-means clustering

The elbow method was first used to determine the optimal number of clusters. As in Figure 8:

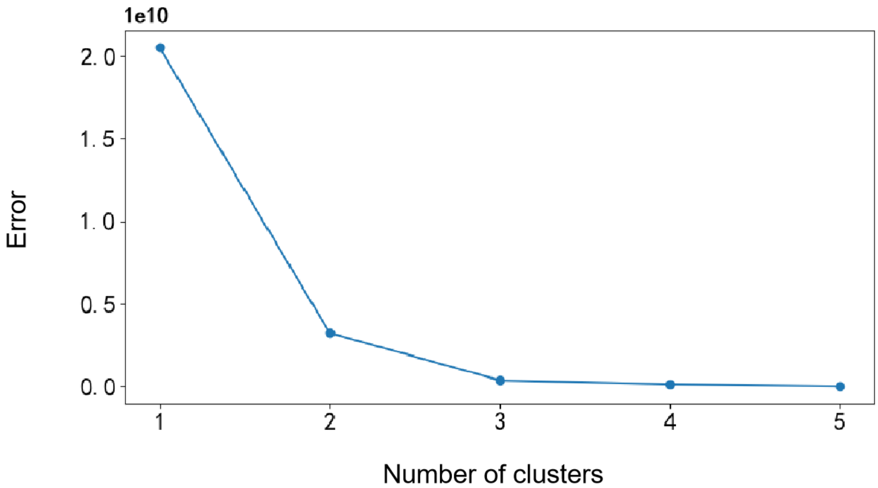

Figure 8. Elbow chart

The optimal inflection point of the elbow diagram is defined as a sharp change in the sum of squares of errors before that point and a smaller change after that point. So, in this paper, 3 is selected as the inflection point, i.e., the clustering is 3 classes. Then K-means clustering is carried out based on sales volume and recommended selling price [8], and labels are added to the data, which is shown in Table 2.

Table 2. Recommended selling prices of individual products

Category name	Sales(kg)	Recommended price	Tags
Aquatic Rootstocks	40581.353	9.689851	1
Foliage	198520.978	6.317257	2
Cauliflower	41766.451	9.138538	1
Solanaceae	22431.782	8.695459	1
Pepper	91588.629	10.578216	3
Edible mushrooms	76086.725	12.036941	3

The first category is a combination of aquatic flowering stems, cauliflowers and eggplants, which has a relatively low sales volume and a medium recommended selling price. The second category is the floribunda category, which has a very high sales volume and a relatively low recommended selling price. The third category is a combination of Chili and Edible Mushrooms, which has a medium sales volume and a relatively high recommended selling price.

Carefully analyzing the relationship between total sales volume and cost-plus pricing (recommended selling price) for each category, this paper finds that the floral and leafy category, which is the second category alone, may need to be restocked more frequently by supermarkets due to its high sales volume and low recommended selling price. The combination of aquatic flower stems, cauliflower and eggplant as the first category will have relatively lower sales volume due to its higher recommended selling price, and the superstore may need more promotional activities to increase sales.

3.2. Designing daily replenishment totals and pricing strategies for each category for the coming week

(1) Data pre-processing

In order to more accurately predict the daily sales volume of each category from July 1 to 7, 2023, we screened the data of June and July from 2020-2022, as well as the data of June 2023 for prediction. Daily sales volume was calculated based on category and date grouping, and some of the data is shown in Table 3 below:

Table 3. Table of single sales by category

No.	Classification name	Date of sale	Single sale(kg)
0	Aquatic rootstocks	2020/7/1	4.85
1	Aquatic rootstocks	2020/7/2	4.6
2	Aquatic rootstocks	2020/7/3	9.572
3	Aquatic rootstocks	2020/7/4	5.439
4	Aquatic rootstocks	2020/7/5	4.019
...	...	...	...
1093	Edible mushrooms	2023/6/26	39.582
1094	Edible mushrooms	2023/6/27	38.708
1095	Edible mushrooms	2023/6/28	53.742
1096	Edible mushrooms	2023/6/29	48.314
1097	Edible mushrooms	2023/6/30	39.572

(2) Time series analysis

Based on the daily sales volume in June and July within the last two years obtained in the previous step, this paper uses the ARIMA model in time series forecasting [9] to predict the daily sales volume of each category in the coming week.

The basic idea of the ARIMA model is that the data series of the forecast object over time is regarded as a stochastic sequence, and a certain mathematical model is used to approximate the description of this sequence. Once this model is recognized it is possible to predict future values from past and present values of the time series.

We use a line graph to visualize the predicted daily sales for each category from July 1 to 7, 2023, as shown in Figure 9:

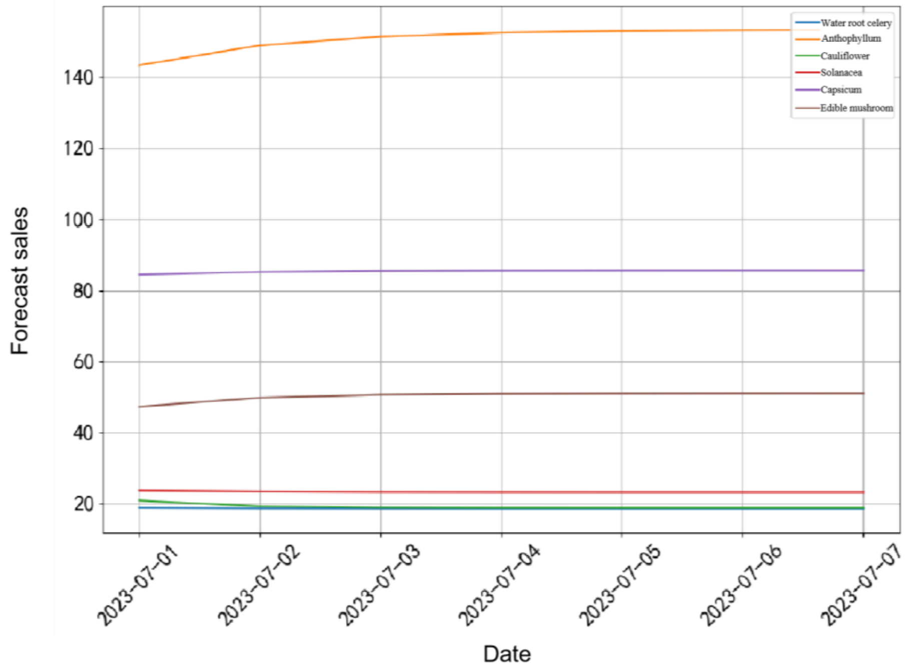

Figure 9. Line graph of projected daily sales for category

As can be seen from the figure, the predicted daily sales volume of these six categories over a 7-day period is basically stable, with the flower and leaf category being the best seller, with daily sales volume greater than 140 kg in all cases, which may be related to consumers' health consciousness. The chili pepper category is the second best-selling category, with a daily sales volume close to 85 kg, which is related to the importance of chili peppers in Chinese cooking. The daily sales volume of edible mushrooms is at a high level and shows an increasing trend, which indicates that consumers recognize the nutritional value of edible mushrooms. The other three categories had poor sales in July as most of the time of listing was concentrated in February to May and the temperature was high, making them not easy to preserve and having a poor taste.

The daily replenishment volume is decided by the superstore in the early morning, which should be understood as the daily purchase volume, i.e. the daily sales volume predicted now. The specific results are shown in Table 4 below:

Table 4. Forecasted daily sales by category

Catagory	7.1	7.2	7.3	7.4	7.5	7.6	7.7
Aquatic Rootstocks	18.8251	18.6748	18.6368	18.6272	18.6247	18.6241	18.6240
Foliage	143.2651	148.901	151.3826	152.4751	152.9562	153.1679	153.2612
Cauliflower	20.9149	19.2962	18.9309	18.8485	18.8298	18.8256	18.8247
Solanaceae	23.6938	23.4152	23.3224	23.2914	23.2811	23.2777	23.2765
Pepper	84.4230	85.1965	85.4764	85.5777	85.6144	85.6276	85.6324
Edible mushrooms	47.3228	49.8335	50.6468	50.9102	50.9956	51.0232	51.0322

(3) Differential evolutionary algorithm

The differential optimization algorithm [10] optimizes the selling price step by step by iterating until the best-selling price that maximizes the profit is found, and the fine-tuning step of the selling price can be adjusted during the iterative process to avoid falling into the local optimal solution.

In this problem, the total cost as well as the predicted sales volume have all been derived earlier, and we multiply the two to get the predicted total cost for each category, which will provide us with an estimate for determining the total cost that each category may incur in the coming week. The results are shown in Table 5 below:

Table 5. Projected total costs by category

Catagory	Projected total cost
Aquatic Rootstocks	956.8001654
Foliage	4572.804312
Cauliflower	884.3281236
Solanaceae	965.6711586
Pepper	4343.426
Edible mushrooms	3007.658834

Also check the highest and lowest selling price of each category and the average value to facilitate the subsequent calculation, the results are shown in Table 6 below:

Table 6. Statistics of selling price of each category

Category name	Minimum selling price	Maximum selling price	Average selling price
Aquatic Rootstocks	1	53.8	9.689851
Foliage	0.1	119.9	6.317257
Cauliflower	0.5	19.8	9.138538
Solanaceae	1.5	21.6	8.695459
Pepper	0.9	79.8	10.578216
Edible mushrooms	0.5	116	12.036941

Firstly, using the calculated highest, lowest and average selling price of each category as data, python is used to create an empty dictionary to store the price boundaries of the categories, with the upper and lower bounds being the highest and lowest selling price respectively. Then seasonality is taken into account, as July is a hot period, so the sales price of each category is fine-tuned. For example, the seasonality factor for the chili pepper category is adjusted to 0.9, and the seasonality factors for the other five categories are increased slightly by 10% or 20%. Then we obtain the predicted sales volume and total cost per unit, calculate the expected revenue, introduce the differential evolution algorithm to find the optimal, and finally obtain the optimized selling price. The optimal selling price for each category is shown in Table 7 below:

Table 7. Optimal selling price for each category

Category	Best price
Aquatic Rootstocks	39.300838
Foliage	24.674987
Cauliflower	16.034739
Solanaceae	3.574877
Pepper	26.582579
Edible mushrooms	37.652743

(4) Calculation of projected reenue

The daily sales volume for each category was explicitly predicted earlier, so it is sufficient to calculate the projected revenue based on the projected sales volume, the projected total cost per unit, and the projected selling price. The formula is as follows:

$$Q = \sum_{j=1}^{6} (PS - PB) \times PX \tag{5}$$

Where Q denotes the expected revenue, j denotes the number of categories, PS denotes the expected selling price, PB denotes the expected total cost per unit, and PX denotes the expected sales volume.

The maximum benefit is found to be \$73, 525.45 by python programming.

4. Conclusion

Fresh produce superstores need to develop a reasonable replenishment plan to cope with the pricing and supply/demand challenges caused by the short freshness period, many varieties and different origins of vegetables. This paper is based on ARIMA-differential evolutionary algorithm to calculate the selling price of vegetable goods and predict the sales.

In the actual vegetable sales and selling price will be affected by many factors, so we should be more comprehensive introduction of other vegetable sales and selling price of the factors affecting the vegetable sales and selling price, which is conducive to vegetable sales and selling price of the model to get a high degree of accuracy and to make reasonable and scientific decision-making.

References

[1] YANG Yue, FENG Gu Cheng, LI Xin Ye, et al. Internet+" sales model of celery products [J]. Changjiang Vegetable, 2020 (20): 4.

[2] LI Zongyou, ZHAO Xiaoli, MA Yinyi, et al. Research on automatic pricing and replenishment decision problem for vegetable commodities [J]. Advances in Applied Mathematics, 2023, 12 (11): 14.

[3] A Qi. Research on Pricing of Scientific and Technological Research Services Based on Cost-plus Pricing Method [J]. Library Research and Work, 2021 (10): 8.

[4] Lv Wang. Research on trend prediction of fresh vegetables sales based on improved SVM [J]. [2024-03-11].

[5] Zhang Yanliang, Liu Ya. Research on fresh e-commerce product sales prediction based on fractional order gray model [J]. Preservation and Processing, 2021, 21 (10): 6.

[6] WU Jiuniu, GAO Decheng, JIANG Weidong, et al. Application of interpolation method based on box-and-line diagram in data processing of empty box barometer [J]. Industrial Instrumentation and Automation Device, 2023 (3): 93 - 98.

[7] ZHANG Huifeng, JING Tian. The combination of cost plus and consumer perceived value: A pricing model for knowledge payment products [J]. Journalism and Communication Research, 2021 (1): 38 - 51, 127.

[8] Zhu J, Wang H. An Improved K-means Clustering Algorithm [J]. Microcomputer Information, 2010, 10 (1): 193 - 199.

[9] Tao Z F, Tan W F, Chen H Y. A time series prediction method for component data incorporating fuzzy time series analysis [J]. Systems Engineering Theory and Practice, 2023, 43 (5): 1534 - 1544.

[10] Ge Yuanyuan, Chen Depo, Zou Feng. Multi-cluster multi-strategy differential large-scale multi-objective optimization algorithm [J]. Control and Decision Making, 2024 (002): 039.