Research on Superstore Vegetable Pricing and Replenishment Prediction Based on ARIMA Time Series and Particle Swarm Algorithm

Xiaoxiao Chang*

School of Economics and Management, Xi'an Shiyou University, Xi'an, China

*Corresponding author: changxiaoxiao331@163.com

Abstract. With the development of digitalization, it is important to provide accurate and reasonable pricing and replenishment strategies for superstores through market demand and data information. In this paper, we first preprocess the relevant data and draw line graphs to compare the correlation between multiple individual items through Spearman correlation analysis. Secondly, Random Forest feature importance analysis is used to explore the relationship between total sales and cost-plus pricing of each vegetable category, and finally, the target data are tested for grade ratio and a gray model is constructed, and model prediction is carried out by using spsspro, and then the particle swarm algorithm is used to derive the optimal pricing strategy. Meanwhile, by updating the pricing regularly and making accurate replenishment decisions, we can better adapt to market fluctuations, obtain better sales opportunities, and maximize the benefits gained by the superstore itself.

Keywords: Spearman's correlation coefficient; random forest; ARIMA time series; particle swarm algorithm.

1. Introduction

Vegetable-based goods in fresh produce superstores tend to have a short shelf life and fluctuate widely in price and demand changes. Purchasing, storage and transportation costs of vegetable products are all important parts of business operations. Good automated pricing and replenishment decisions can help optimize vegetable sales and inventory management and improve supply chain efficiency. So in order to maximize the benefits of superstores, it is necessary to give accurate pricing and replenishment strategies. In this paper, firstly, we analyze the pattern and correlation between the sales of each vegetable category and single product and conclude that the sales of flower and leafy vegetables are much higher than other categories. Then the relationship between the total sales of vegetable categories and cost-plus pricing is explored, and it is concluded that the importance of cauliflower features is the greatest. Finally, a particle swarm optimization algorithm model is established to further solve the pricing strategy, so that the planning of pricing and replenishment strategy is more accurate, and the superstore's own interests are maximized.

2. Relationship between each vegetable category and sales of individual items

In order to more intuitively analyze the relationship between commodities and the sales volume of superstores, this paper collects three years of data on the sales volume of each vegetable in a large superstore. Firstly, the data are processed in terms of category and sales date, and then the time-sales line graphs of monthly sales volume of each category, quarterly sales volume of each category, and quarterly sales volume of the top 20 individual items are plotted to explore the pattern between sales volumes. When exploring the correlation, a linearity test and a positivity test are first performed to see if the data fit, and if the fit is good, Pearson's correlation coefficient analysis is used; conversely, Spearman's correlation coefficient analysis is used. Here the distribution of monthly sales per category is plotted (see Fig. 1), and the distribution of quarterly sales per category is plotted (see Fig. 2).

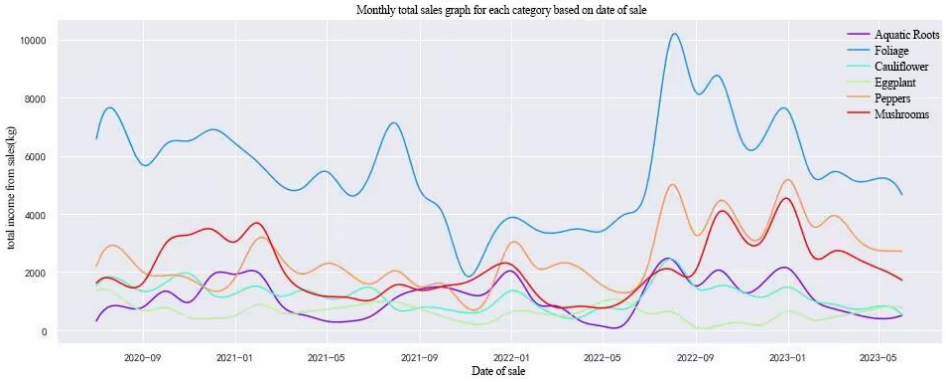

Fig. 1 Distribution of monthly sales volume per category of goods

Looking at Fig. 1, it appears that overall, the fluctuations in the increase and decrease in the sales volume of each type of vegetable show a broadly consistent trend. There are two nodes where the changes are more pronounced. At the end of 2021, sales of all types of vegetables were at their lowest in recent years due to a more severe epidemic, and starting in May 2022, as the weather turned warmer and the epidemic eased, sales of each vegetable category increased, reaching a new high in sales in August and September. Although the trend of changes in each category is similar, there is a clear difference in sales. Flowering and leafy vegetables have always occupied the top of the sales list in the vegetable category, with the exception of the end of 2021, when sales were low due to the epidemic when they were in the leading position in terms of sales. In addition, sales of eggplant vegetables fluctuated but were lower with or without the epidemic. Sales of aquatic root and cauliflower vegetables were slightly more than those of eggplant; sales of edible mushrooms were higher than those of chili peppers before the seriousness of the epidemic, and sales of chili peppers were higher than those of edible mushrooms after the epidemic and remained so for a long time. It is possible that the epidemic increased people's health awareness and reduced their consumption of edible mushroom vegetables.

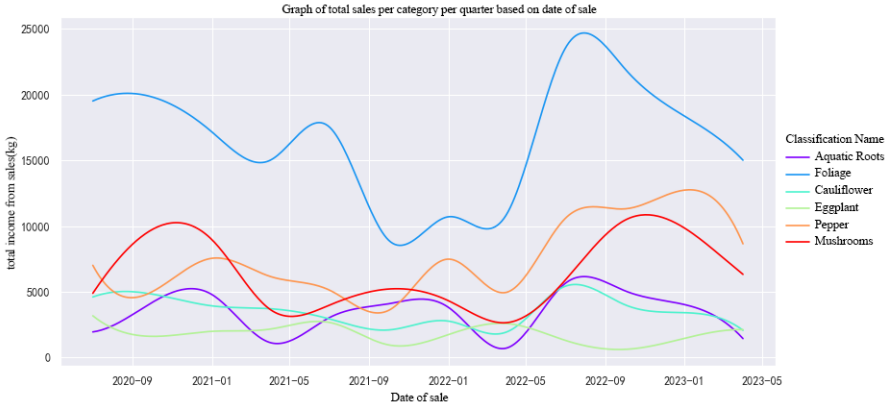

Fig. 2 Distribution of sales per category per quarter

Observing Fig. 2, changing the interval of change from monthly to quarterly, the distribution pattern of each category is further obvious, and again, the sales of flowering and leafy vegetables are still much higher than other categories, and the sales of eggplant vegetables are the lowest overall.

Based on the analysis above, the Spearman correlation coefficient is utilized here for calculation, and the Spearman correlation coefficient is applicable to data that do not satisfy linear correlation or data whose distribution does not satisfy normal distribution. Therefore using Spearman correlation analysis, correlation is calculated by converting the data into an order (rank) [1].

$$\rho = 1 - \frac{6\sum d_i^2}{n\left(n^2 - 1\right)} \tag{1}$$

Where d_i denotes the difference between the rank values of the ith data pair and n denotes the total number of observed samples.

Firstly, the correlation analysis of the single day sales of each category is performed to get the correlation heat map as shown in 3. The results are observed to explore the correlation.

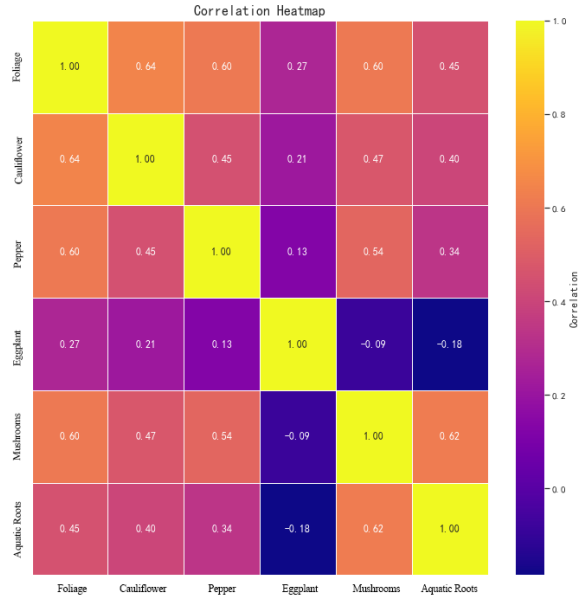

Fig. 3 Heat map of correlation of daily vegetable sales by category

Observing Fig. 3, the daily sales of eggplant vegetables are less correlated with other categories of vegetables and have a weak negative correlation with edible mushrooms and aquatic root vegetables. Whereas, the daily sales of cauliflower and leafy vegetables have a higher correlation with all other categories of vegetables (except eggplant). Further observation of the results yields a higher correlation between sales of cauliflower and leafy vegetables and a higher correlation between sales of edible mushrooms and aquatic root vegetables. The categories with higher sales correlation can be sold in adjacent positions to increase vegetable sales. At the same time, the correlation analysis of the daily sales of each individual item is then performed using the same method. A correlation heat map is obtained as shown in Fig. 4, observe the image and explore the correlation.

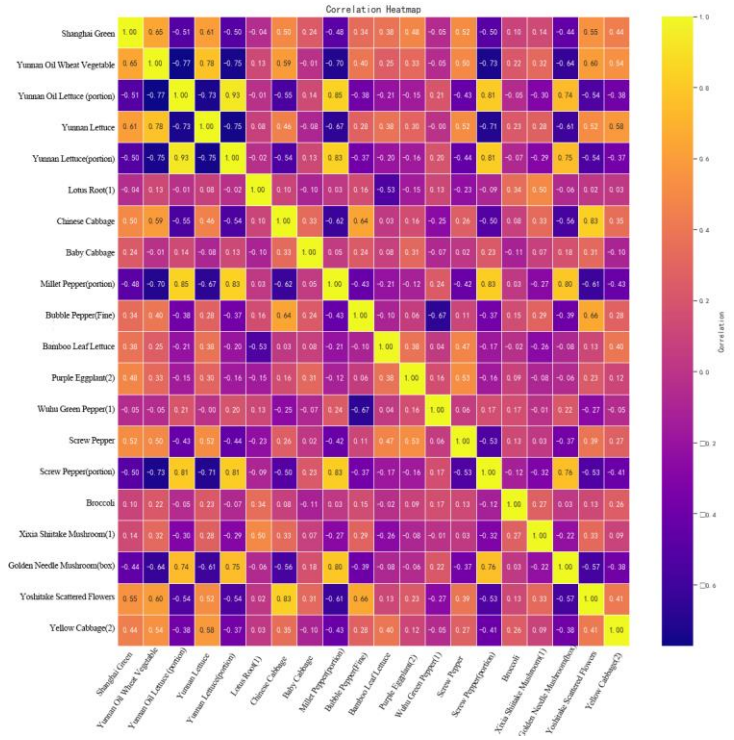

Fig. 4 Heat Map of Correlation of Daily Sales of Vegetables by Individual Product

Observing Fig. 4, it is found that there is a strong correlation between these single items, which is a good inspiration for merchants. You can adjust the position of the higher correlation of the single product, placed in the adjacent position, to improve sales, similar to the idea of placing beer next to the mother and baby products. You can also carry out a combination of marketing buy and give way, that is, carry out the purchase of high-volume vegetables to give away low-volume vegetables to drive sales. Or bundled sales, the implementation of discounts and other preferential policies to improve sales.

3. Exploring the relationship between total sales and cost-plus pricing

In this paper, the relationship between total sales volume and cost-plus pricing of each vegetable category is investigated by means of Random Forest Feature Importance Analysis (RFIA). Random forests are used to evaluate the feature importance, i.e., to quantify the contribution of each feature to the classification performance of the constructed ll decision trees [2].

Define the indicator function.

$$I(x,y)=\begin{cases}1, & x=y\\ 0, & x\neq y\end{cases} \tag{2}$$

The m feature F_m in the kth tree of $FIM_{\mathrm{lm}}^{((\mathrm{O}))}$ is the

$$FIM_{\mathrm{km}}^{(\mathrm{OOB})}=\frac{\sum_{p=1}^{n_0^k} I\left(Y_p,Y_p^k\right)}{n_0^k}-\frac{\sum_{p=1}^{n_0^k} I\left(Y_p,Y_{p,\pi_m}^k\right)}{n_0^k} \tag{3}$$

When feature F_m does not appear in the kth tree, the $FIM_{\mathrm{lm}}^{(00\mathbf{B})}=0$.

The importance score of feature F_m in the whole random forest is defined as

$$FIM_{\mathrm{m}}^{(\mathrm{OOB})} = \frac{\sum_{k=1}^{K} FIM_{hm}^{(\mathrm{OOB})}}{K\sigma} \tag{4}$$

Where: k denotes the number of decision trees in the random forest; σ denotes the standard deviation of $FIM_{km}^{(00B)}$. The importance score F_m of feature $FIM_m^{(OOB)}$ characterizes the contribution of F_m to the correct classification rate.

This results in the characteristic importance of each category as shown in Fig. 5:

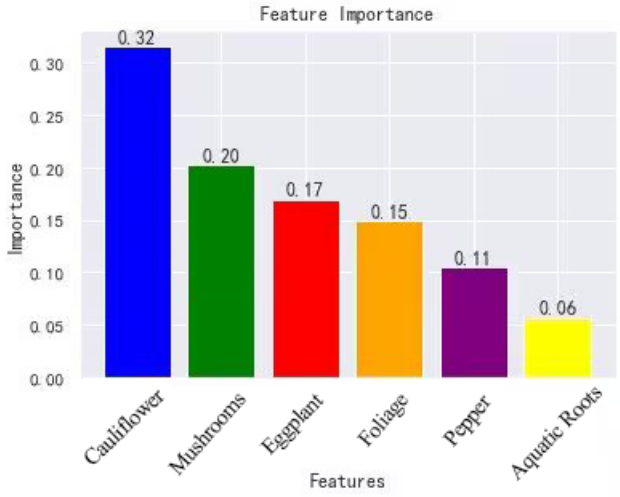

Fig. 5 Importance of features

Observing Fig. 5, we can see that cauliflower has the greatest importance of characteristics, while edible mushrooms, eggplant, foliage, chili peppers, and aquatic rhizomes have the lowest importance of characteristics. It can be seen that in all vegetable categories, the consumer's purchase behavior of cauliflower category is more affected by the price, more sensitive to the price, for this type of product, subsequent pricing can start from the price.

4. Total replenishment explored

In order to be more intuitive, the total daily sales of each category were visualized and analyzed using Excel to obtain the data graph shown in Fig. 6. Due to the large amount of data, only the total sales of cauliflower category is shown in this paper.

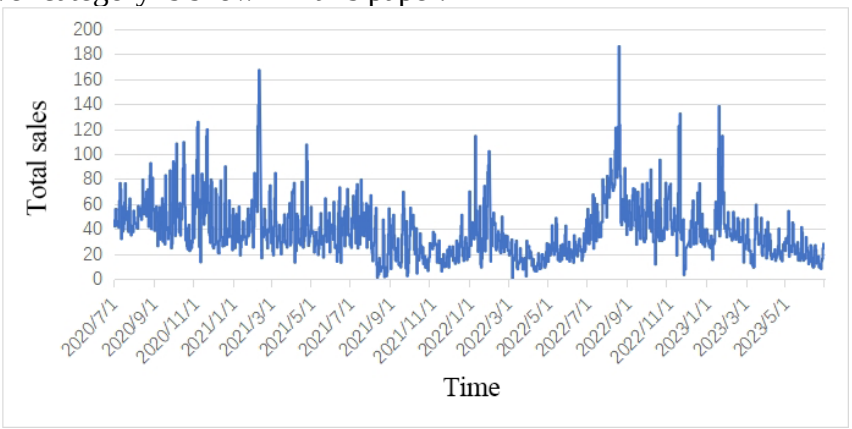

Fig. 6 Total daily sales of cauliflower vegetables

As can be seen from Fig. 6, this time series does not have seasonal cyclical variations, so in this paper, we consider the use of ARIMA time series model to forecast the daily replenishment of each category of superstores from July 1-7, 2023 The model is solved using spsspro [3], resulting in the time series forecasting plot shown in Fig. 7:

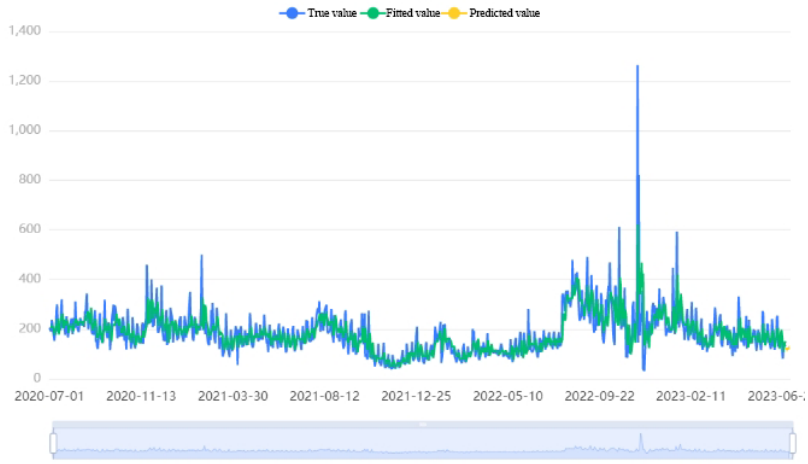

Fig. 7 Time series prediction of flowering and leafy vegetables

The final daily replenishment for each category for July 1-7, 2023 is shown in Table 1:

Table 1. Forecast of daily replenishment of vegetables by category

Time	Cauliflower	Foliage	Peppers	Eggplant	Mushroom	Aquatic Roots
2023/7/1	22.86313342	125.3744893	79.63786317	22.65976715	53.04192677	20.08432817
2023/7/2	20.42336895	116.1213273	77.06739626	20.83798178	42.0327373	20.99868672
2023/7/3	19.2755183	118.8441391	74.60735471	17.68379978	32.23793608	21.3721996
2023/7/4	18.72715764	121.3903454	73.67191006	17.32198486	43.52417894	21.73601746
2023/7/5	18.45697981	125.1676275	77.26180417	16.67265377	51.61525805	22.09039194
2023/7/6	18.31588786	126.5047565	79.04300537	18.39614063	44.55870522	22.43556816
2023/7/7	18.23469595	125.4959971	79.22547396	19.39791819	45.45473751	22.77178487

5. Pricing Strategy Exploration

In order to develop a more accurate pricing strategy, it is necessary to consider the wholesale price of each category, the average wastage rate, and the expected sales volume. We will use a "cost-plus pricing" strategy that takes into account the cost of the item and the expected sales volume to determine an appropriate selling price [4].

The specific steps are as follows:

Step1: Combine the wholesale price, the wastage rate and the predicted sales volume to calculate the cost of each category. The cost (per kilogram) of each vegetable category was obtained as shown in Table 2.

$$Cost\ per\ kg = retail\ price \times (1 + attrition\ rate) \tag{5}$$

Step2: Determine the markup rate based on expected sales volume.

150% markup if expected sales are <20kg (because these are low-volume items and want to get a higher return per unit).

If the expected sales volume is < 50 kg, the mark-up rate = 130% (medium volume goods). If the expected sales volume is higher, mark-up = 120% (high volume goods, where it is hoped to increase the sales volume through a lower price).

Step3: Calculate the sales price for each category.

$$Sales\ price\ per\ kg = cost\ per\ kg \times markup\ rate \tag{6}$$

Finally, the following result was obtained by python programming:

Table 2. Seven-day pricing for each vegetable category

Time	Cauliflower	Foliage	Peppers	Eggplant	Mushroom	Aquatic Roots
2023/7/1	9.96	6.89	9.3	6.78	8.75	16.12
2023/7/2	9.96	6.89	9.88	6.78	8.75	16.12
2023/7/3	9.96	6.89	9.88	6.78	8.75	16.12
2023/7/4	9.96	6.89	9.88	6.78	8.75	14.68
2023/7/5	9.96	6.89	9.88	6.78	8.75	14.68
2023/7/6	9.96	6.89	9.88	6.78	8.75	14.68
2023/7/7	9.96	6.89	9.88	6.78	8.75	14.68

6. Modeling of single-item replenishment volume scheme based on gray prediction

From firstly this paper preprocesses the data to get 27 sequences of single product sales volume, and uses spsspro to carry out the level ratio test, and secondly all the level ratio values of the level transformed sequences are located in the interval (0.779, 1.284), which indicates that the level transformed sequences are suitable for the construction of the gray prediction model [5].

For the sequence of sales volume of single product that does not pass the level ratio test, the sequence is "shifted and transformed", so that the sequence after shifting and transforming meets the level ratio test. Using spsspro to solve the model, the a posteriori difference ratio of this sequence is $0.125 < 0.35$, which indicates that the model is highly accurate. And finally, the restocking quantity data of these 27 individual products on July 1, 2023 is obtained.

7. Modeling of pricing strategy scheme based on particle swarm algorithm

The particle swarm algorithm optimization model is built using the replenishment quantity of a single product and each constraint as required in the previous question [6].

The quantity ordered for each individual item satisfies the minimum display quantity constraint, i.e., it can be expressed as the total quantity of each individual item M_i, $i = 1, 2, 3 \ldots .27$ needs to be greater than 2.5kg, which can be expressed as

$$M_i..2.5 \tag{7}$$

For the calculation of profit, we use the cost-plus pricing established in the second question and set W_{i} to be the profit margin. Thus, the pricing for each individual product can be expressed as

$$P_i = K_i\left(1 + C_i\right) \tag{8}$$

where P_i denotes the pricing of the ith individual item, K_i denotes the cost price of the ith individual item, and C_i is the attrition rate of the ith individual item. Therefore, the final total profit is expressed as:

$$W_{\text{total}} = \sum_{i=1}^{27} \frac{M_i k_i C_j}{100} \tag{9}$$

In summary, the optimization model constructed for this problem is:

$$\max W_{\text{total}} = \sum_{i=1}^{27} \frac{M_i k_i C_j}{100} \tag{10}$$

$$\text{s.t.} \begin{cases} P_i = K_i\left(1 + C_i\%\right) \\ M_{\text{i}}..2,5 \\ 0 < C_i < 100 \\ k_i, P_{\text{i}} \in R \\ i = 1,2,3,4\cdots 27 \end{cases} \tag{11}$$

During the iteration process, each particle updates its own position and velocity based on the individual historical optimal position and the group historical optimal position as well as its own position and velocity until the stopping condition is satisfied, and finally obtains the optimal pricing of the 27 individual products.

8. Summary

In this study, the replenishment and pricing problems in fresh vegetables sales are effectively solved through the comprehensive application of data integration, Spearman correlation analysis and gray prediction. With data integration through Python's pandas library, this paper not only ensures data integrity, but also simplifies the preprocessing process. In finding the optimal pricing strategy, the particle swarm algorithm is used to improve the computational efficiency, and although there are limitations in dealing with extreme cases and the assumption of linear relationship, the model as a whole shows strong practicality and potential for generalization, which is applicable to not only the sale of vegetables, but also to the pricing and replenishment strategy of other commodities. In the objective optimization of replenishment and pricing, the particle swarm algorithm, which has the advantages of strong global search capability and fast convergence speed, is used, which can lead to a more accurate and globally optimal solution in a shorter period of time. By adopting this algorithm, we are able to optimize the replenishment and pricing strategies more effectively and improve the efficiency and accuracy of decision making.

References

[1] LI Yuan,LIU Yutian,FENG Liwei. Nonlinear dynamic process feature extraction and fault detection based on Spearman correlation analysis[J]. Journal of Shandong University of Science and Technology(Natural Science Edition),2023,42(02):98-107.DOI:10.16452/j.cnki.sdkjzk.2023.02.011.

[2] H.R. Yang,Y.C. Chen,Anna Zhao et al. Construction of a diagnostic model for peri-implantitis based on random forest and artificial neural network[J/OL]. West China Journal of Stomatology,1-13[2024-02-05].http://kns.cnki.net/kcms/detail/51.1169.R.20240201.1437.002.html.

[3] LIN Yanquan,LI Ming,TANG Wan et al. Time series analysis and prediction of the number of reported cases of adverse drug reactions in China based on ARIMA model[J]. Pharmacy and Clinical Research, 2023,31(06):519-522.DOI:10.13664/j.cnki.pcr.2023.06.006.

[4] An Q., Wang S. Y., Du Yidong et al. Micro-credit technology:Developmental changes,internal logic and future outlook--Based on the perspective of loan cost-plus pricing[J]. Qinghai Finance,2020,(09):33-40.

[5] Lin Yuhan. Research on new energy automobile enterprise value assessment based on gray prediction[J]. Modern Industrial Economy and Informatization,2023,13(12):252-255.DOI:10.16525/j.cnki.14-1362/n.2023.12.081.

[6] Y.B. Zhang,C.L. Xiang,W.D. Wang et al. Coordinated control strategy for model-predicted torque of distributed electric drive vehicles based on particle swarm optimization-ant colony fusion algorithm[J]. Journal of Military Engineering,2023,44(11):3253-3268.