Research on automatic pricing and replenishment decision of vegetable commodities based on penalty function LSTM model

Honglin Liao\*,#

Fuzhou University

Fuzhou, China

13055971609@163.com

Jiachen Huang#

Fuzhou University

Fuzhou, China

jiacheng.huang@ieee.org

Shujuan Chen#

Fuzhou University

Fuzhou, China

[whihhcau@163.com](mailto:whihhcau@163.com)

#These authors contributed equally to this work

*Abstract*—This paper mainly proposes an optimization model for supermarket vegetable sales. By establishing the distribution law of sales volume, the relationship between category and item sales volume, and the fitting model of the relationship between sales volume and cost plus pricing, the paper aims to make replenishment and pricing decisions for supermarket, so as to achieve better operating earnings. The graph of past sales volume is drawn and the cyclical and seasonal trends of sales volume are found. Taking the month as the classification unit, the distribution relationship of the sales volume of each category and single product with the month is established. The trend of merchandise sales and the impact of holidays on sales are analyzed. Spearman coefficient analysis was used to study the correlation between category and single product, and the combination of category and single product with strong correlation was found. The weighted average method was used to evaluate the different contribution degree of each item to its category, and the daily selling price of each item was calculated. Based on the reasonable assumption, the profit model of supermarket is established, and the sales volume and markup rate are positively correlated with the profit of supermarket. The penalty function -LSTM model is used to fit the relationship between sales volume and selling price to realize the pricing strategy of price maximization. Forecast the daily replenishment volume and sales price from July 1 to 7. In general, this paper comprehensively uses statistical analysis, weighted average, penalty function -LSTM model and other methods to solve the problem of supermarket vegetable sales, and provides relevant decision-making suggestions and strategies.

Keywords—spearman correlation coefficient, weighted average, Replenishment pricing strategy, Penalty function -LSTM model

# Introduction

Due to the short shelf life of vegetable commodities and the deterioration of products with the increase of sales time, the supermarket should be replenished every day according to the historical sales and demand of the goods to ensure the freshness and supply of the goods. However, since the purchase and transaction time of vegetables is usually in the early morning, businesses need to make replenishment decisions of various vegetable categories on the day when they are unsure of the single items that can be purchased on the day and their wholesale prices. In terms of sales, merchants use the "cost plus pricing" method to price vegetables, and the goods with poor transport damage and product phase change are usually sold at a discount. Therefore, reliable market demand analysis is essential for replenishment and pricing strategies. Market demand analysis can be seen from the demand side and the supply side. From the demand side, there is a certain correlation between the sales volume of vegetable commodities and time. From the supply side, the supply variety of vegetables is more abundant from April to October. Therefore, many scholars have conducted research on it.

Ding Haifeng et al. [1] analyzed the trend of vegetable planting innovation in China. In addition, due to the short shelf life of vegetables, the sale cycle of vegetables is too long, which will increase the preservation cost of vegetables. Cheng Jiahui [2] analyzed the impact of the length of vegetable sales cycle on vegetable prices. Qiao Xue [3] studied the joint order pricing strategy aiming at quantity loss and quality loss to reduce the adverse effects of product corruption and uncertain demand in the retail process of fresh products. Zhang Jinlong et al. [4] established a joint decision model of perishable new product pricing and dynamic batch replenishment, and designed a solution algorithm. The results show that the joint decision of pricing and replenishment is necessary. The optimal price has a U-shaped relationship with product diffusion speed and repurchase rate, which decreases first and then increases. The optimal price increases with the increase of consumers' willingness to buy, and decreases with the increase of product spoilability. Based on the analysis of the characteristics of mass customization, Wei Zee et al. [5] put forward a cost-plus pricing method suitable for mass customization, and analyzed the advantages and disadvantages of this method. Cost-plus pricing means that the price of the product should cover the cost of production and sales and get a reasonable return [6]. Gu Sihong [7] built a dynamic pricing model of fresh products considering the change of freshness, and aimed to maximize profit to obtain the optimal pricing of fresh products in the freshness period and the freshness decline period, and analyzed and studied the results of the model.

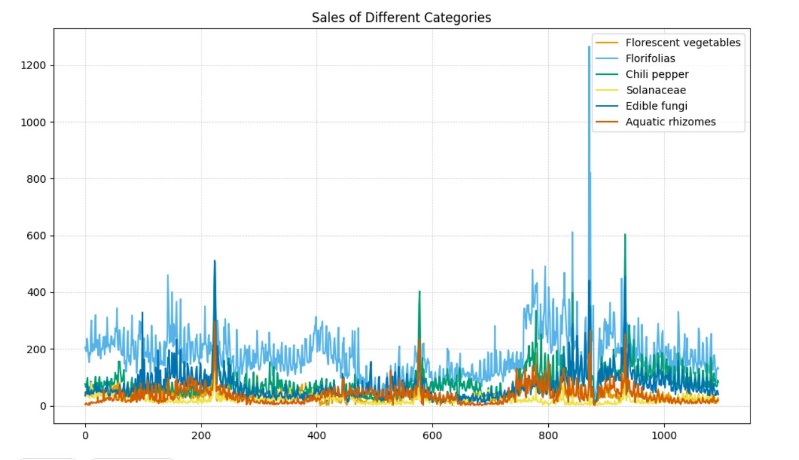
The purpose of this paper is to make replenishment and pricing decision for supermarket to maximize the profit, which has important practical and theoretical significance, and provides strong support and guidance for supermarket vegetable sales management and decision. Through in-depth analysis of sales data and establishment of sales models, business super can more accurately grasp the sales situation, formulate reasonable replenishment and pricing strategies, so as to improve business efficiency. The application of the model is helpful to reduce the rate of commodity loss, reduce excess inventory and loss through reasonable replenishment decisions, and improve the efficiency of resource utilization. To sum up, this study provides feasible management strategies in actual operation, and also plays a certain role in innovation and promotion in research methods and theory, which is of great significance to the development of supermarket vegetable sales management and related fields.

# Model building and solving

## Distribution law analysis of sales

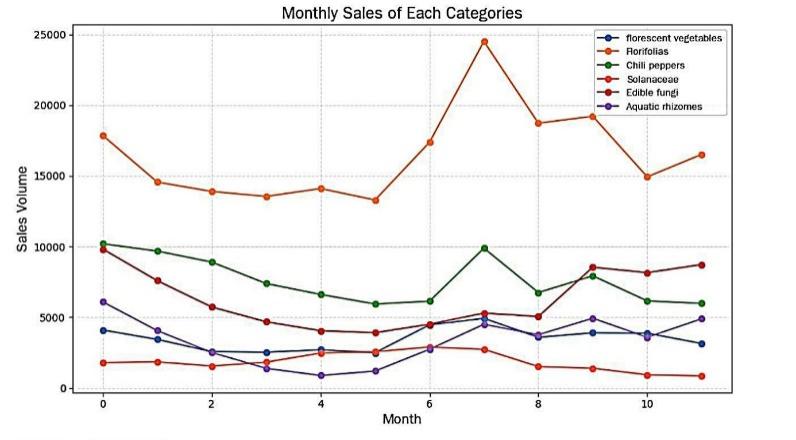
The data in this paper is a total of 878,504 sales detailed data from July 1, 2020 to June 30, 2023. Due to the large amount of data, it is necessary to preprocess the data in order to prevent the error of the data from affecting the calculation result. The data is pre-processed, such as eliminating outliers, filling in blank values for smoothing processing, and aggregating the sales volume of each item in each category.

After pre-processing, relatively accurate sales data are obtained, and the sales volume of each category in each month of each year is obtained by analyzing the data, as shown in Figure 1. It can be seen that the curve presents a large cyclical and seasonal, the change trend is roughly the same. Considering that a closer timeline is helpful to accurately grasp the distribution law, this paper chooses the month as the classification unit to establish the distribution relationship between the sales volume of each category and each single product with the month.



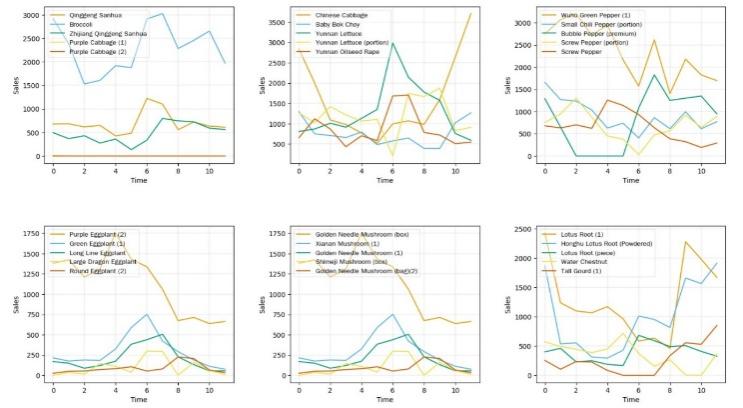
1. Sales volume of all sales details

The sales volume of each category in each month of each year is added to obtain the total sales volume of this category in each month, and then the distribution relationship of the sales volume of this category is drawn on a monthly time scale, as shown in Figure 2. It can be seen that from the overall trend, the sales volume of all categories will start to fall from January, and then gradually show an upward trend after June, and gradually decline after the sales volume reaches the peak at the end of August. Sales of edible mushrooms peaked at the end of October and began to gradually decline. It is speculated that the relationship between festivals leads to similar trends in the sales of various categories. For example, during the Spring Festival, the sales of various dishes will increase significantly. In terms of quantity proportion, the Mosaic category contains more single varieties, and the overall sales volume is higher than other categories, reaching the highest in August, which can be close to 25,000 kg; The sales of nightshade accounted for the least in the overall sales, and the monthly sales were stable at less than 2500 kg.



1. Monthly sales volume for each category

In this paper, the sales volume of each item in each month of each year is added up to obtain the total sales volume of each item in each month. Due to the large amount of data, this paper selects the distribution relationship of the sales volume of the top 5 items in each category, as shown in Figure 3.



1. The monthly sales volume of the top 5 items in each category

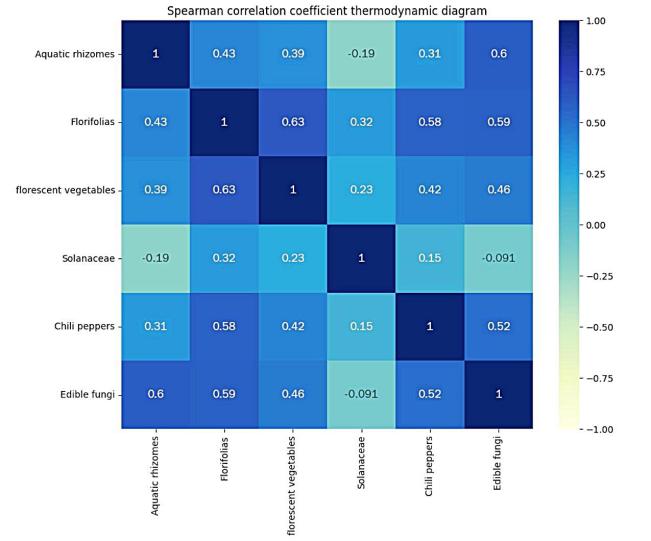
## Correlation evaluation model based on Spearman coefficient analysis

From the perspective of category and item, all data have nonlinear relationship and sequential characteristics. Therefore, Spearman coefficient analysis is chosen in this paper to understand the relationship between the two dimensions of category and item.

Spearman's coefficient analysis is a non-parametric statistical method that is suitable for measuring a monotone relationship between two variables, whether linear or non-linear. The results correctly assess the degree of correlation between the two variables. Unlike Pearson's correlation coefficient, Spearman's coefficient does not require the variables to follow a normal distribution. The calculation formula of Spearman's coefficient is shown in equation (1), where and are the positions of x and y respectively; ,are the average positions of x and y, respectively.



The basic process is based on the sorted variables, assign a rank to the value of each variable, calculate and square the rank difference of each variable, and substitute the calculation formula of Spearman coefficient to obtain Spearman coefficient. The value of the coefficient ranges from -1 to 1, where -1 indicates complete negative correlation, 0 indicates no correlation, and 1 indicates complete positive correlation, which is used as the evaluation basis for the degree of correlation between the two variables.



1. Thermal maps of pairwise correlations of various types

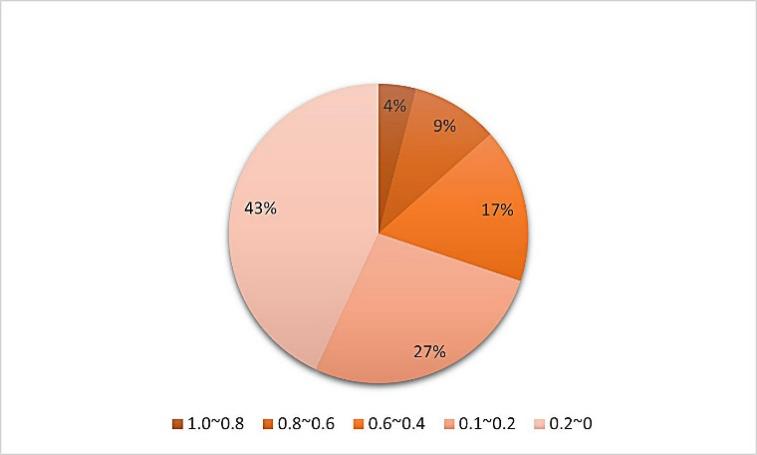
Figure 4 shows the heat map of pairwise correlation between each category. The depth of each cell color in the heat map indicates the strength of the correlation between the corresponding categories. A lighter color or a value near 0 indicates a weak correlation, while a darker color or a value near 1 or -1 indicates a strong correlation. By observing the heat map, it is not difficult to see that the top 4 pairs with the strongest correlation and the top 4 pairs with the weakest correlation are shown in Table 1:

TABLE I The correlation between the two categories is top 4

|  |  |
| --- | --- |
| Top four strongest correlations | The top four weakest correlations |
| Cauliflower - Flowers and leaves | Nightshade - Edible fungi |
| Edible fungi - aquatic rhizomes | Nightshade. - Capsicum |
| Edible fungi - flowers and leaves | Aquatic roots - Nightshade |

We can see the strongest correlation between cauliflower and Mosaic, probably because they are similar in flavor and taste, and consumers have a higher tendency to buy them at the same time. The association between nightshade and edible mushroom was weakest, possibly because nightshade and edible mushroom may not have much overlap in their culinary uses. Nightshade is usually used for stir-frying, barbecuing, or soups, while edible fungi are usually used for stir-frying, souping, or cooking hot pot. Therefore, at the time of purchase, consumers may be more inclined to choose one according to their culinary needs, rather than buying both products, resulting in a weak correlation between them.

Since the number of items is 251, the amount of data to calculate the pduo relationship between 251 data is very large. The strength of the correlation is also not easy to accurately judge, especially when the correlation is very close to 1. For example, after ranking by correlation strength, even in the first 200 pairs, the correlation is still 1. Therefore, when looking at the data, you cannot simply select the first few that have the strongest correlation. In order to solve this problem, the following method is adopted in this paper: the correlation coefficient range of [0,1] is divided into 5 intervals, each of which has a step size of 0.2. The closer the absolute value of the correlation coefficient is to 1, the stronger the correlation is. The closer to 0, the weaker the correlation. Then, the pie chart is used to visually display the distribution of correlation strength, as shown in Figure 5.



1. Proportion chart of correlation between pairings of an item

By observing Figure 5, it can be found that the proportion of high correlation is 4.05%, the proportion of high correlation is 9.40%, the proportion of moderate correlation is 16.70%, the proportion of weak correlation is 26.75%, and the proportion of weakest correlation is 43.15%. From the overall point of view, the correlation between the sales volume of 251 items is not very strong, which may provide certain guidance and decision-making basis for subsequent sales strategy, inventory management, market analysis and other aspects.

## Ridge regression fitting relationship

After data preprocessing, the single product purchased in each transaction record information from June 1, 2020 to June 30, 2023 and the corresponding category of the single product are obtained, and the relevant data is calculated, as shown in the following equation.















Formula (2) is the relationship between the selling price, wholesale price and markup rate of item j in item i. Formula (3) is the relationship between the wholesale quantity, sales quantity and loss rate of each item of item j in category i; Formula (4) adds the daily total sales volume of the same category to get the daily total sales volume of 6 categories; In formula (5), represents the daily markup rate of item i, represents the daily selling price of item i, and represents the daily wholesale price of item i.

In order to calculate the selling price and wholesale price of each category more accurately, this paper takes into account the extent to which each item in each category contributes to the overall selling price and wholesale price of the category. Therefore, this paper takes days as the statistical time unit to carry out a weighted average of the selling price and wholesale price of each item in the same category. The purpose of weighted average is to assign different weights to different items in calculating the sales price and wholesale price of the category according to the proportion of each item to the sales volume of the category. Therefore, the daily sales price of item i is shown in equation (6). Where Sĳt stands for the daily sales quantity of item j of category i, Si stands for the daily sales quantity of item i, and Pĳt stands for the daily selling price of item j of category i.

The sale price of the same item may vary from day to day due to different sales times. In order to calculate the daily sales price of each item more accurately, the different sales price of the same item every day is converted into the daily sales price of the same item by using the method of average value Pĳt. Therefore, the daily wholesale price calculation formula of category i is shown in equation (7). Where, Nĳt represents the daily wholesale quantity of the JTH item of category i, Ci represents the daily wholesale quantity of the JTH item of category i, and Cĳt represents the daily wholesale price of the JTH item of category i. The conversion formula between Sijt and Nijt is derived by equation (3), as shown in equation (8). From this, the daily wholesale quantity of the JTH item in category i is derived, and Ci is obtained by summarizing.

Limited by the sample data, Ridge regression method is used to fit the relationship between the sales quantity and the markup rate of each vegetable category.In the process of solving the model, a ridge parameter λ is defined to control the intensity of the regularization term. The ridge regression model is established according to the sample data, and the regression coefficient is solved by minimizing the objective function. According to the regression coefficient obtained, the fitting model is constructed, and the prediction and analysis are carried out.

## Future replenishment pricing strategy based on LSTM

First, determine the wholesale price Ci of each category for the next 7 days and obtain the wholesale price Ci and attitudeLI of each category for each day from July 1, 2021 to June 30, 2023. In order to determine the wholesale price of each category in the next 7 days, taking into account the impact of rest days and working days on the market, as well as the seasonality of commodity supply, 8 complete weeks are selected, and it is observed that the standard deviation of each category is not large during this period. Therefore, this paper decides to choose the average wholesale price of each category in this period as the wholesale price of the corresponding category in the future from July 1 to 7. By calculating the average wholesale price of each category in this period, a relatively stable reference value can be obtained as the forecast of the wholesale price of the corresponding category in the future from July 1 to 7. Considering the impact of rest days and working days on the market, as well as the seasonality of commodity supply, this paper chooses a longer period of time for calculation in order to improve the reliability of calculation results.

Then, determine the attrition rate for each category over the next 7 days. According to the attached data, it can be observed that the loss rate is only related to the single product and has nothing to do with time. Therefore, in order to determine the attrition rate of each category in the next 7 days, this paper decides to use the weighted average of the attrition rate of each category as the attrition rate of each category. Considering that the loss rate is independent of time, this paper selects the loss rate of each category as the basis, and calculates the average value to represent the loss rate of the whole category. The rationality of this method is that by averaging the loss rate of each product category, the volatility of data is reduced and the accuracy of prediction is improved.

Finally, the scheme is determined to realize the influence of past sales on future sales and the possible complex nonlinear relationship between sales quantity and pricing in the sales data of vegetable categories with the maximum supermarket income. In order to guide the model to learn more about the relationship between sales volume and pricing in order to maximize supermarket returns, this paper adopts a method based on LSTM time series model compound penalty function to predict sales volume and pricing in the next week. Since the cost plus pricing method is suitable for quick formulation of pricing strategy, easy to control, applicable to different products and stable, this paper chooses the cost plus pricing method to formulate the pricing strategy for vegetable categories. A regression linear equation based on sales volume and pricing fitting was added to the loss function of LSTM model training as a penalty function to guide the model to learn and balance the relationship between sales volume and mean pricing to maximize the quotient surplus return, and a specific objective function was constructed. Based on equation (2) and equation (3), the objective function of quotient surplus return was given in this paper, as shown in the following equation.



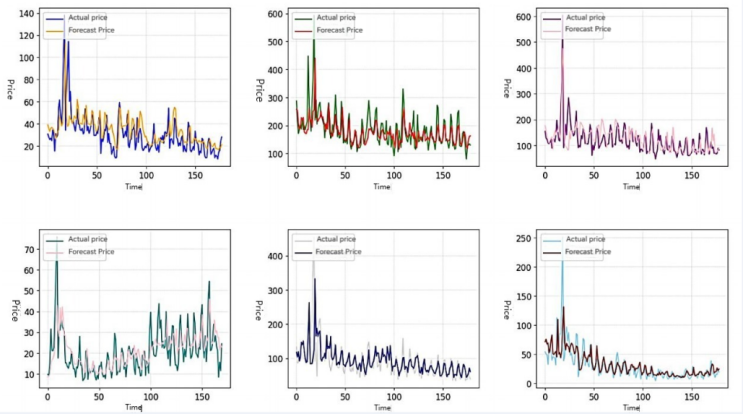
Take the sales volume of each category as the base column, shift the base column down by 7 units, and create 7 new columns representing the sales volume of the past seven days respectively, so as to build a time series data set that enables the LSTM model to learn the sales pattern and trend for seven days. Each row features sales for the past seven days, and the label is sales for the next day. 80% of the data is used as a training set to train the model, and 20% is used as a test set to test the model performance.

By observing the function relationship of excess revenue constructed above, when the commodity loss rate and wholesale price are relatively fixed, the commodity sales volume and mark-up rate have a positive correlation with excess revenue. By using the previously fitted regression function on commodity sales volume and commodity markup rate as a penalty function, and compounding with mean square error, the model is constrained and guided to learn the characteristics of the correlation between commodity sales volume and commodity markup rate, so as to maximize the return of supermarket.

LSTM model construction: In the first layer, a one-dimensional convolution layer is constructed to extract local features of the input data. By setting the weight of the school filter, different features of different positions of the data can be captured, which plays a significant role in learning local patterns and trends of the LSTM timing model. It then builds a BatchNormalization layer that helps speed up training and improve model generalization by batch normalizing the output of the previous layer so that the mean and variance of each feature are close to 0 and 1. After this, two LSTM layers are built. The LSTM layer can be modeled from sequence data to capture long-term dependencies in the sequence. At the same time, add three Dropout layers, and randomly discard 40%,30%, and 20% data respectively to their Settings to prevent overfitting.

The LSTM model is trained with the pre-prepared time series data, and the parameters of the model are updated by backpropagation and optimization algorithm, so that the model gradually converges. Evaluation metrics such as mean absolute error (MAPE) are used to evaluate the performance of the model on the test set, and the model is optimized to improve the model's expectations. The trained LSTM model is used to predict the future time series, and the relevant indexes are calculated to evaluate its performance.

According to the previous literature [7], dynamic pricing should be used to maximize the return of supermarket. Appropriately raise pricing when sales are high, and appropriately lower pricing when sales are low. Obtain the daily replenishment quantity and price increase rate of each category that can make the supermarket benefit the most in the next 7 days. By substituting the data, the model MAPE is 0.2306, which has good adaptability. The forecast results are as follows:



1. The fitting curve of sales volume and cost plus pricing for each category

# SUMMARY

The model constructed in this paper preprocesses the data before use, including replacing outliers and finding missing values. Ensure that the data smoothing quality of the model is high, thus improving the reliability of the model. Understanding competitors' pricing strategies and promotional activities can help supermarkets refer to competitors' market positioning and strategies. The supermarket can determine its own pricing strategy according to the pricing strategy of its competitors in order to remain competitive. In addition, competitor data can also help merchants understand the degree of price competition in the market, so as to adjust their replenishment volume and pricing strategy to better cope with market competition. Conducting market research can help supermarkets understand the overall needs and trends in the market. Market research can collect data through questionnaires, interviews, observations and other means to understand consumers' purchase intentions, consumption habits, price sensitivity and other information. Through the analysis of market research data, the supermarket can better understand the direction of vegetable seed industry organization and business model change, so as to adjust pricing strategies and improve market competitiveness. The collection of customer feedback data can help supermarket to understand customer preferences, preferences and changes in demand for vegetable products. By analyzing customer feedback data, supermarket can understand which vegetable categories are favored by customers and which categories have room for improvement. Based on customer feedback data, the retailer can adjust replenishment volume and pricing strategy to better meet customer needs. However, the model in this paper uses linear fitting and cannot capture the nonlinear relationship in the relationship: in fact, the relationship between sales volume and cost plus pricing is not necessarily linear, and this paper treats it as linear, ignoring the influence of marginal effect. This may lead to poor fitting of the model to nonlinear relationships and failure to accurately capture complex relationships between variables. To better fit the relationship between volume sold and cost-plus pricing and capture the complex relationship between the two, it may be necessary to consider using other non-linear models or improving the network structure.

# References

1. Ding Hai-Feng, YU Muan-Cang, Wang De-Xin, et al. Analysis on innovation trend of vegetable seed industry in China [J]. China Vegetables, 2015 (8): 1-7. (in Chinese)
2. Cheng J H. Research on the impact of vegetable sales cycle on vegetable market price [J]. China Collective Economy,2015(4):85-86. (in Chinese) DOI:10.3969/j.issn.1008-1283.2015.04.044.
3. Qiao X. Joint replenishment pricing strategy of fresh products considering sales loss [D]. Jiangsu: Southeast University,2021. (in Chinese)
4. Zhang Jinlong, WU Xiang, XU Haoxuan. Perishable products pricing and replenishment joint decision model [J]. Journal of systems engineering, 2018 (1) : 79-89. The DOI: 10.13383 / j.carol carroll nki. The jse. 2018.01.008.
5. Wei Ze E, Chen Gang, Ding Shengchun, et al. Research on cost plus cost pricing method of large gauge moulded products [J]. Heilongjiang Foreign Economy and Trade,2007(2):84-85.
6. Han Junhua, Kam Seung Dao. Review of cost-plus pricing [J]. Finance and Accounting Monthly (Accounting Edition),2012(8):74-75.
7. Gu Sihong. Research on dynamic pricing of fresh products at H retailer considering freshness change [D]. East China university, 2023. DOI: 10.27012 /, dc nki. Gdhuu. 2023.001318.