Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 12, No. 1, 2024

Research on Vegetable Sales Volume and Pricing Strategy Based on GWO-LSTM Modeling

Zihan Wang [1, *, †], Xinyi Liu [2, †], Yuzhuo Wang [1, †]

[1] School of Mechanical Engineering, Qinghai University, Xining, China
[2] School of Finance and Economics, Qinghai University, Xining, China
[*] **Corresponding author**: Zihan Wang (Email: 2624097401@qq.com)
[†] These authors also contributed equally to this work

Abstract: This paper focuses on an in-depth study of the problem of sales volume and pricing strategy for vegetable categories, aiming at constructing an automatic pricing and replenishment model to maximize revenue. First, by analyzing the sales data of each category and single item of vegetables, the mean value of sales volume is calculated, and the correlation between each category of vegetables is visualized using charts and Pearson correlation coefficient heat maps, and it is found that there is a significant correlation between specific categories. Secondly, this paper adopts the cost-plus pricing method to calculate the markup rate of each individual item and combines it with linear regression analysis to reveal the relationship between markup rate and sales volume, and further predicts the sales volume and price in the coming week through the GWO-LSTM model, which provides data-driven pricing and replenishment strategies for the superstore based on the data. Finally, for the data of a specific time, this paper develops a linear programming model to formulate a replenishment and pricing scheme that meets market demand and maximizes revenue. The methodology and findings of this study have important practical applications for retail formats such as fresh food superstores.

Keywords: Vegetable sales volume; Cost-plus pricing; Pearson correlation coefficient; GWO-LSTM model; Linear programming.

1. Introduction

The short shelf life of goods and the gradual deterioration of their quality over time as they are sold becomes a notable challenge in the current retail industry, especially in supermarket sales of fast-consuming goods such as vegetables and fruits. This phenomenon leads to the risk of vegetable items being difficult to sell the next day, forcing supermarkets to rely on historical sales and demand forecasts to decide on replenishment of goods. Considering the diversity of vegetables, their different origins and irregular delivery times, supermarkets struggle to make accurate replenishment and pricing decisions in the absence of specific information. This study aims to address three main issues: first, to analyze the sales patterns of different varieties and individual vegetables and their interrelationships; second, to explore how to maximize supermarket benefits through replenishment planning for variety units under a fixed weekly replenishment quantity and pricing strategy; finally, considering the limitations of the selling space, it is proposed that under the premise of controlling the number of commodity types, in order to satisfy the market demand while maximizing the benefits of the commodity replenishment volume and pricing for refined planning. Through these analyses, this paper aims to provide more scientific and effective replenishment and pricing strategies for supermarkets to optimize their operating efficiency [1].

2. Relevance Analysis

In conducting the data pre-processing process, this study classified the data into six categories for analysis through thorough data cleaning, missing value processing, and data categorization based on storage duration. The percentage of each category was determined through cluster analysis and basic statistical indicators of sales volume such as mean, and media were calculated using statistical software. The study also identified and removed outliers in the data and discretized the continuous variables to reduce the effect of observation error on the model, thereby reducing the risk of overfitting and improving model stability. Through the lookup function in the spreadsheet software, the accurate matching of vegetable category and individual item code information was realized, and the cost pricing method was applied to calculate the pricing of vegetable commodities. In terms of model building and solving, this study quantitatively analyzed the relationship between vegetable categories, individual product categories and sales volume, and revealed the correlation between categories and individual products through standardization and market distribution observation, as well as scatter plotting and the application of correlation coefficient model. The analysis showed that the average sales volume of chili pepper items was the lowest in different time periods, while the average sales volume of aquatic root and tuber items was the highest, and the sales volume of flower and leaf items had the greatest fluctuation with seasonal changes, and these findings provided a scientific basis for the development of effective replenishment and pricing strategies as shown in Figure 1.

As can be seen from Figure 1 above, the share of the cauliflower category is relatively stable, the share of the foliar category and pepper category is steadily increasing year by year, aquatic rhizome category shows a trend of decreasing firstly and then increasing, and edible fungi as a whole shows the trend of increasing, but there is a sharp decline in the share of the phenomenon. The Flower and leaf category accounted for the most up to 37.8%, the pepper category accounted for 23.7%; the eggplant category accounted for the least, only

5.1%, and the aquatic rhizome category accounted for 6.7%. The line graph is shown in Figure 2 and the distribution matrix in Figure 3.

This study explored the potential correlation between vegetable category and unit sales volume by applying correlation analysis such as Figure 4, specifically Pearson's correlation coefficient. After standardizing the data to eliminate the effect of magnitude and to ensure normal distribution of the data, the correlation analysis revealed a significant positive correlation between category and unit sales volume. Pearson's correlation coefficient is close to 1 ($\rho = 0.34$**) and the significance level is less than 0.001, which means that the linear relationship between the two sets of variables is extremely significant. Through hypothesis testing, the original hypothesis (H_0: $\rho = 0$) was rejected and the alternative hypothesis (H_1: $\rho \neq 0$) was supported, which is that there is a non-zero correlation between sales volume and broad category. Moreover, the 95% confidence interval (0.032, 0.036) further confirms the stability of the correlation [2].

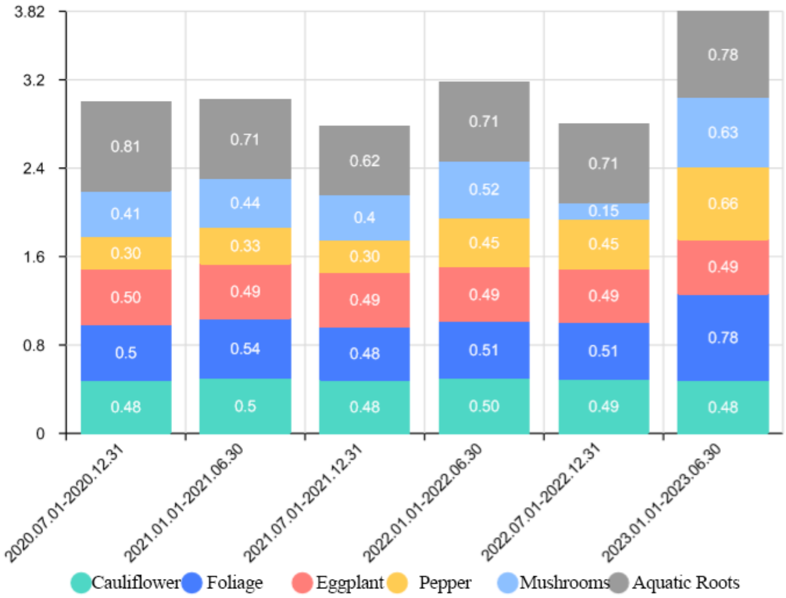

Figure 1. Histogram of the percentage share of each major vegetable category over time

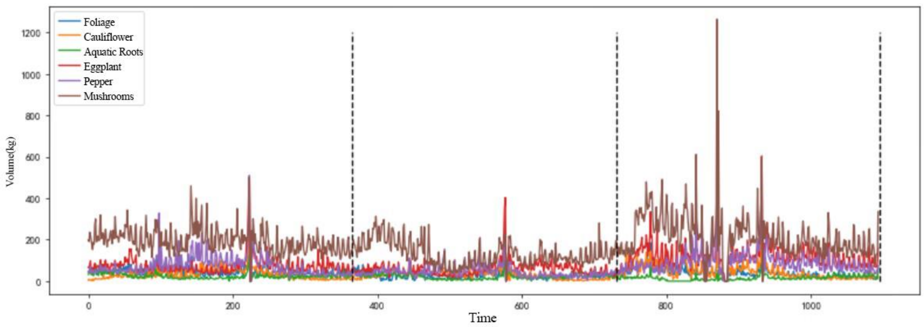

Figure 2. Line graph of sales volume over time by broad vegetable category

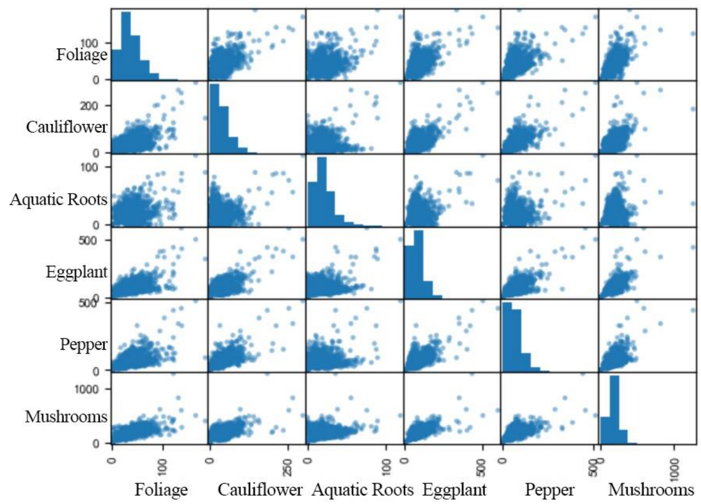

Figure 3. Dispersion matrix

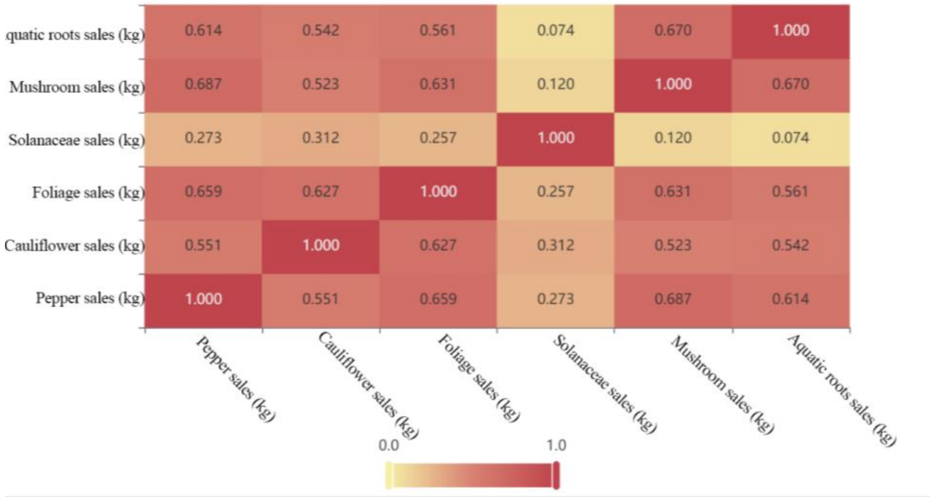

Figure 4. Heat map of correlation coefficients

This study reveals the correlation between the sales volume of different vegetable categories through the application of heat map four, where the change in the shade of color visually indicates the magnitude of the values. The results of the analysis showed that the correlation between the sales of edibles and chili peppers was high, while the correlation between the sales of eggplant and aquatic roots as well as edibles was low. In addition, the results of the cluster analysis showed the frequency of the different cluster categories and their percentage of the total, where cluster category one accounted for the majority (92.683%) and the rest of the categories accounted for a smaller percentage.

The significance analysis further explored the variability of the two variables, sales volume, and item code, between the different clustering categories. For the sales volume variable, its p-value of 0.174 is higher than the conventional significance level of 0.05, indicating that there is no significant difference in sales volume between the different clustering categories and hence the original hypothesis of no difference cannot be rejected. For the item coding variable, its significance P-value is 0.000, which is much lower than the threshold of 0.05, indicating that there is a significant difference in item coding between different clustering categories, thus rejecting the original hypothesis. The F-test results show that there is a significant difference in the variance of the item coding variable, implying that the correlation between different individual items is stronger, which provides an important statistical provides an important statistical basis for supermarkets' category-specific sales strategies and inventory management.

3. Category pricing and replenishment strategies

3.1. Modeling

In this study, a comprehensive forecasting and optimization framework is constructed by combining the cost-plus pricing method and linear regression analysis, as well as the Long Short-Term Memory (LSTM) neural network and Gray Wolf Optimization (GWO) model. By reviewing the relevant literature and applying the cost-plus pricing method, the pricing strategy of the goods was determined and the relationship between total sales and cost pricing was revealed by linear regression analysis. This relationship is used as a constraint for the LSTM model to predict the daily sales volume for the coming week. Considering the possible loss of goods during transportation and sales, this study used the GWO model to optimize the replenishment and pricing strategies aiming to maximize the total revenue of the supermarket. By combining the cost-plus pricing method with the calculation of gross sales, costs, and operating profit, as well as the calculation of replenishment volume after considering the wastage rate, this study provides a comprehensive revenue growth strategy for supermarkets [3].

LSTM neurons have "forgetting gates", "input gates" and "output gates" and two cellular candidate states. The LSTM cell state under this structure can transfer information over time and complete the deletion of information through the three gates, which makes the LSTM have a memory function. The structure is shown in Figure 5.

First, Oblivion Gate performs information screening to eliminate some of the information with the following formula.

$$f_t = Sigmoid\left(W_f\left[X_t, h_{t-1}\right] + b_f\right) \tag{1}$$

Were. f_t is the forgetting gate matrix of the neuron at the t time; and W_f is the oblivion gate weight matrix; and X_t is the input matrix of the neuron at the moment of t the input matrix of the neuron at the moment; and h_{t-1} is the hidden state output matrix of the neuron at time $t-1$; and b_f is the oblivion gate bias matrix.

Second, the input gate determines which information about the cell state is retained, with the following expression.

$$I_t = Sigmoid\left(W_I\left[X_t, h_{t-1}\right] + b_I\right) \tag{2}$$

$$C_t = Sigmoid\left(W_c\left[X_t, h_{t-1}\right] + b_c\right) \tag{3}$$

$$C_t = f_t \odot C_{t-1} + I_t \odot C_t \tag{4}$$

where: I_t is the input gate matrix of the neuron at time t; C_t is the cell candidate state matrix of the neuron at time t; is the cell state matrix of the neuron at time t; W_I is the input gate weight matrix; b_I is the input gate bias matrix; W_c is the cell candidate state weight matrix; b_c is the cell candidate state bias matrix; and $\odot$ is the dot product of the matrices.

Finally, the output gate determines the output of the neuron at the current moment with the following expression.

$$o_t = Sigmoid\left(W_o\left[X_t, h_{t-1}\right] + b_o\right) \tag{5}$$

$$h_t = o_t \odot \tanh C_t \tag{6}$$

Were. o_t is the$t$ the output gate matrix of the neuron now of time; W_o is the output gate weight matrix; and b_o is the output gate bias matrix; and h_t is the t the hidden state output matrix of the momentary neuron.

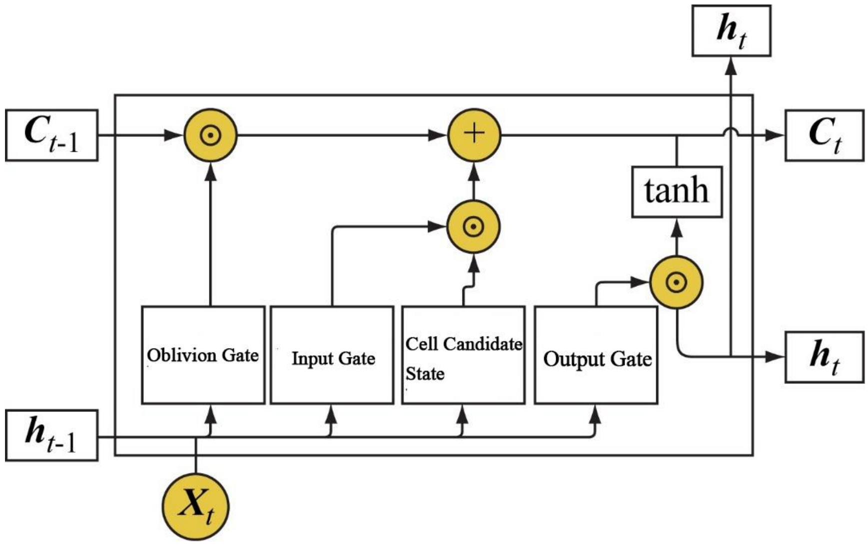

Figure 5. LSTM structure

The Gray Wolf Optimization (GWO) algorithm is a group intelligence optimization algorithm that simulates the hunting behavior of gray wolves, and it searches for the optimal solution by simulating the social hierarchical structure of wolves and their hunting strategies. The GWO algorithm divides the wolves into four classes: Alpha (leader), Beta (deputy leader), Delta (subordinate members), and Omega (the lowest class), in which the Alpha wolf is responsible for guiding the whole wolf pack [4].

The hunting process of the GWO algorithm is divided into three main phases: encircle, hunt and attack. In the encirclement phase, the wolves surround the prey under the leadership of the Alpha wolf, and the algorithm simulates this behavior by adjusting the position vector of the gray wolf. In the hunting phase, the Alpha, Beta, and Delta wolves are usually closest to the prey due to their position in the pack, so the pack approaches the prey led by these three wolves. The position of the Omega wolf is updated based on the position of the Alpha, Beta, and Delta wolves. Finally, in the attack phase, the optimal solution is obtained by decreasing the control parameters to simulate the process of the wolves approaching and capturing the prey.

The core of the GWO algorithm lies in simulating the social behavior and hunting strategy of wolves and using the hierarchical structure and collective wisdom of wolves to guide the search process to find the optimal solution. The effectiveness of the algorithm is not only reflected in its ability to simulate the hunting behavior of wolf packs in nature, but also in its high efficiency and robustness in dealing with complex optimization problems, which makes it a powerful tool for solving a variety of optimization problems.

3.2. Solving the model

In this study, this paper explores the correlation between markup rate (denoted as y) and sales volume (denoted as x) by constructing a regression model between the total sales volume and total cost pricing of vegetable commodities. Linear analysis was conducted using SPSS software with the aim of revealing the correlation between these two variables. The model assumptions include a disturbance term that is not directly observable that satisfies specific conditions. The results of the analysis showed that the absolute value of the correlation coefficients between the model error term and the dependent variable was small, indicating that they were uncorrelated, thus suggesting that the regression model was well exogenous and had a small error. After eliminating the abnormal data and analyzing the residuals, accurate regression coefficients are obtained in this paper, which provides a solid foundation for the subsequent model building and analysis [5].

Further, in this study, a Long Short-Term Memory (LSTM) network was constructed using MATLAB, including an input layer, multiple hidden layers, a fully connected layer, and an output layer. The data of markup rate and sales volume were fed into the input gates of the network, which were selected and computed by the network, and finally the prediction results were obtained through the forgetting and output gates. The first eighty data points were used for network training and the rest of the data were used for testing and validation, and after 1200 rounds of training and six thousand iterations, the sales prediction results for each category of vegetables were obtained. These prediction results are highly consistent with the original data, the root mean square error (RMSE) of the LSTM network is less than 1 on the training, testing and validation sets, and the R-squared value of the LSTM is almost overlapped with the fitted straight line, which proves the high accuracy and prediction ability of the model. These findings provide dedicated support for sales strategies and cost-plus pricing of vegetable commodities.

For each of the six categories of vegetables, this paper shows the linear relationships with normal p-p plots and the results of GWO-LSTM with scatter plots and line plots. Take the cauliflower category as an example in Figure 6:

In the preliminary analysis, the confidence intervals of the regression coefficients mostly contain the origin less accurately, so this paper transforms the model, performs the residual analysis and eliminates the outliers, The final regression coefficients are. β =-0.019, α = 6.470, y = -0.019x+6.470, at the same time, test the regression model of the three statistics: correlation coefficient R square = 0.063, F = 73.822, and F corresponds to the probability of p = 0.0000, and regression coefficients of the confidence interval of the regression coefficients only a few contain the zero point, indicating that the model regression is more ideal.

Use Deep Lrerning Toolbox data toolbox in matlab to construct LSTM network, which includes input layer, hidden layer, fully connected layer, output layer, hidden layer, the data related to markup rate and sales volume are input into the input gate of the model respectively, and then the neural network is used to select the appropriate value into the forgetting gate, and the output gate is run through the results of the operation of the input gate and the forgetting gate The result of the input gate and the forgetting gate. In this paper, the first eighty data are used for training, and the remaining data are used for testing and validation and 1200 rounds of training, a total of six thousand iterations, the prediction results of each category are shown in Figures 7 and 8.

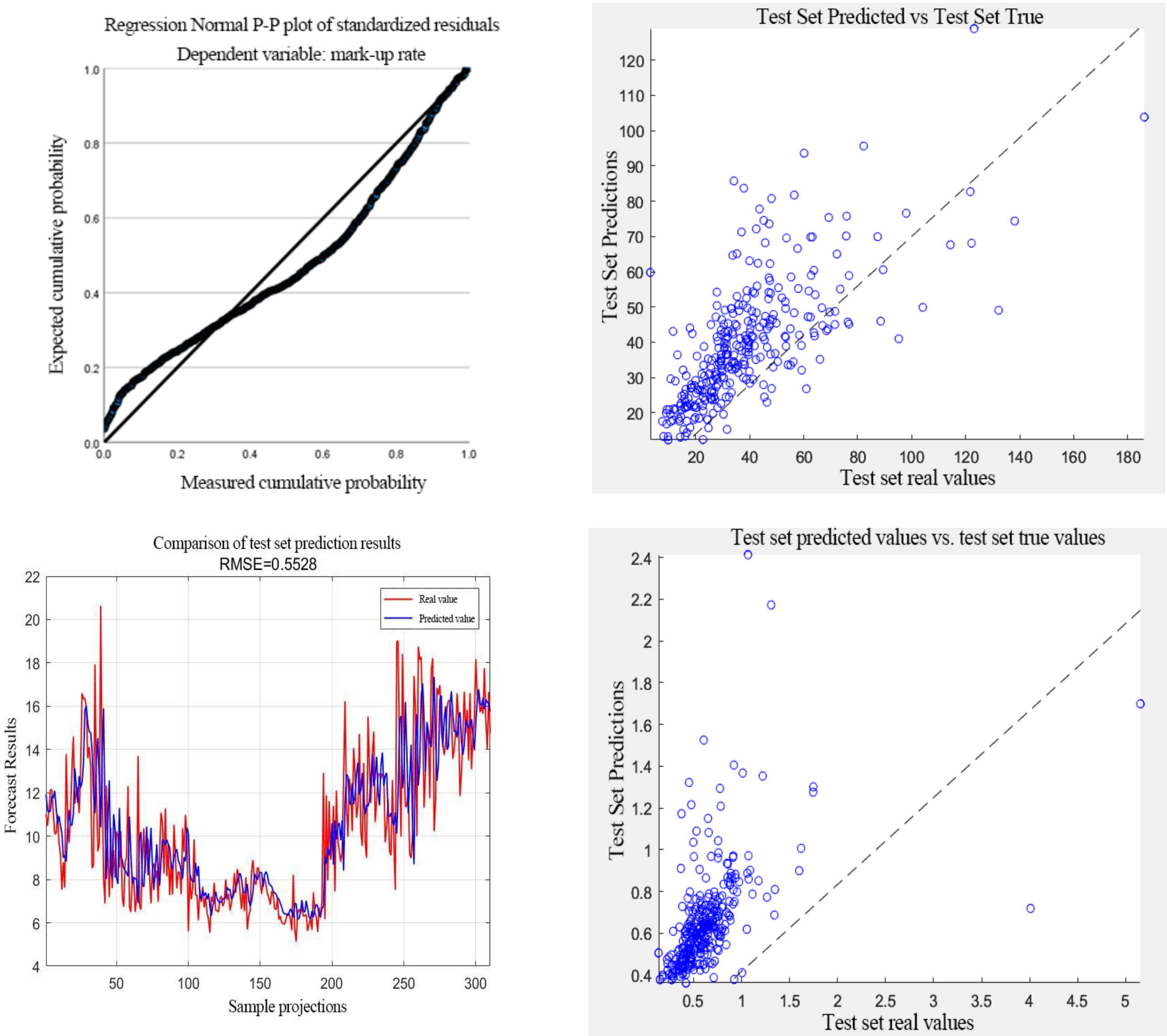

Figure 6. Summary of linear analysis and time series forecasts for cauliflower species

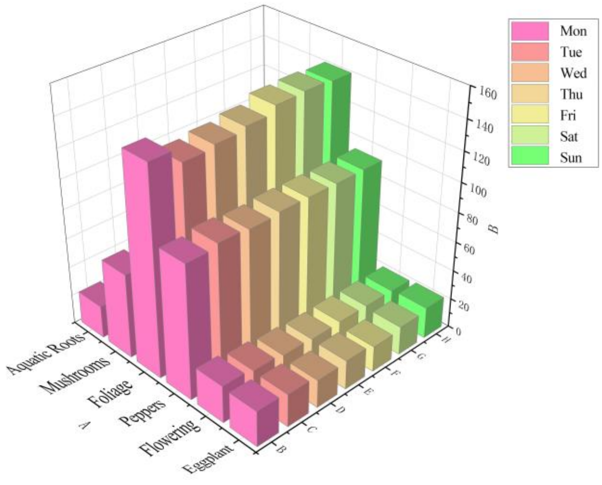

Figure 7. 3D bar chart of trends in replenishment of vegetables by category over a one-week period

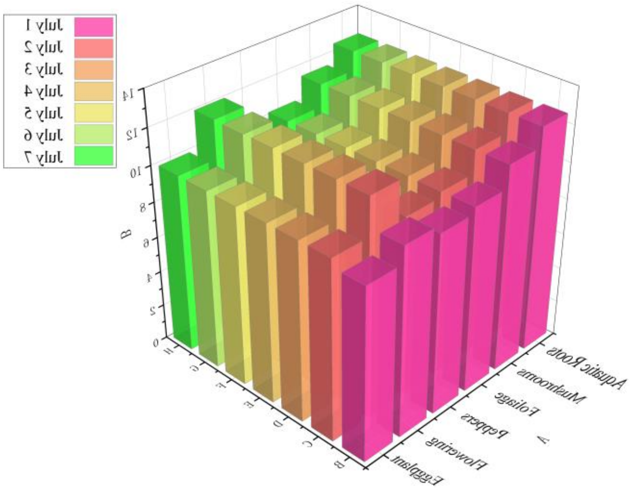

Figure 8. 3D bar chart of trends in optimal pricing of vegetables by category from July 1 to July 7

4. Individual product pricing and replenishment strategies

4.1. Modeling

Extract the sales volume data of vegetable goods and classify the goods according to their individual products to understand the supply and demand and sales of individual products, gross profit = sales volume * (sales price - wholesale price) - sales volume * wastage rate, which requires this paper to give the optimization model for a single day, this paper will use the linear programming model to start from the linear regression relationship found between the daily replenishment and the pricing strategy, and then according to the conditions of Equation 7 to list the corresponding values of the range of values, and further refinement on the basis of this, and finally find the optimal solution.

The simplex method is one of the most used and effective algorithms for solving linear programming problems [6]. A solution that satisfies the constraints is called a feasible solution of a linear programming problem, and the feasible solution that maximizes the objective function is the optimal solution:

$$s.t\begin{cases} max\ P = \sum(p_i) * s_i * y_i - c_i * x_i\,, i \in \{1,2,\cdots,n\} \\ 27 \le y_i \le 33, i \in \{1,2,\cdots,n\} \\ x_i \ge 2.5 * y_i, i \in \{1,2,\cdots,n\} \\ \sum(x_i * b_i) \le \sum(p_i * s_i) \\ \sum(x_i * l_i) \le 0.15\sum(s_i * p_i) \end{cases} \tag{7}$$

4.2. Solving the model

A note on constraints: lety_i Be the total number of sales items, which needs to be controlled between 27 and 33, and let x_i Be the minimum display quantity to satisfy the x_i Greater than 2.5 display requirements, and according to the total sales and cost-plus pricing method of linear analysis of its relationship with the profit listed in the constraints of the objective function. Among them, this paper set p_i Is the selling price of a single product, the s_i Is the replenishment quantity of a single product, and y_i Is the selection variable, and c_i Is the cost incurred in the total process of selling goods, and l_i Is the loss rate in the transportation selling process, andb_i Is the wholesale price of the goods, which needs to be maintained less than the total selling price. Finally, SPSS was utilized to derive the relevant results as shown in Table 1.

Table. 1 Projected result

Daily replenishment	1.0185	0.9716	0.7538	1.1677	2.0756	0.3034
Name	Cucumber	Baby Chinese cabbage	Ice plant	Chrysanthemum coronarium	Cabbage	Purple eggplant
Optimal pricing	6.21	7.52	17.17	22.26	6.74	11.39
Daily replenishment	1.0002	0.3674	0.8472	0.7606	1.2483	0.9118
Name	Wuhu Green Pepper	Yunnan oilseed rape	Chili	Green pepper	Agaricus bisporus	Green & Red Pepper Combo
Optimal pricing	10.00	5.32	17.09	10.62	9.94	4.15

5. Conclusions

In this paper, a prediction model for the relationship between vegetable commodity sales volume and cost pricing is developed by using Pearson's correlation coefficient, linear regression model, long and short-term memory network (LSTM) and gray wolf optimization (GWO) algorithms. The simple operation and easy-to-understand data output of the Pearson correlation coefficient algorithm facilitated the analysis of the correlation between the two variables; the linear regression model transformed the statistical problem into an equation solving problem, which ensured the practicality and wide applicability of the model; the LSTM algorithm showed a strong advantage in dealing with the long-series data and real-time updating, which provided the predictive model with accurate prediction capability; and the GWO algorithm's simple data requirements and ease of implementation make it an effective tool for solving predictive models. However, the model also has certain limitations, such as failing to fully consider the impact of weather changes on vegetable merchandising, as well as efficiency issues and introduction of subjectivity in data processing and algorithm implementation. In the future, optimizing the Gray Wolf Optimization algorithm by introducing methods such as Levy flight, greedy principle and adaptive convergence strategy can further improve the analysis efficiency and accuracy of the model. In addition, by utilizing the advantages of GWO algorithm in processing big data, it can play a key role in the field of network publicity and marketing, helping enterprises to enhance their brand image and attract potential customers, to occupy a favorable position in the competitive market environment.

References

[1] Li Mingyu. Research on the construction and application of sales performance evaluation index system of large-scale commercial super liquor[J]. China Brewing, 2023, 42(10):256-262.

[2] WANG Jinwei, SAN Yuhan, SAN Baohai. Analysis of short track speed skating competition data based on Pearson's correlation coefficient[J]. Ice and Snow Sports, 2023,45(04):9-12+17. DOI: 10. 16741/j.cnki.bxyd.2023.04.002.

[3] Chunxiao Zhang. Optimization method for oil and gas well production prediction based on deep long and short-term memory neural network[J]. Petrochemical Applications, 2023, 42(11):28-31.

[4] Chenxi Zhang. Improved gray wolf optimization algorithm based on inverse tangent inertia weights[J]. Digital Technology and Application,2023,41(12):5-8. DOI:10.19695/j.cnki.cn12-1369. 2023.12.02.

[5] Sun W, Zhou JL. Research on Internet loan default prediction based on oversampling logistic regression model[J]. Journal of North China University of Science and Technology (Social Science Edition), 2024,24(01):54-61.

[6] ZHAO Yanlong,SUN Yaohua,GAO Cailailei et al. Optimization of process parameters of chlorine-containing external drainage water treatment based on simplex method[J]. Yunnan Chemical Industry, 2023,50(10):38-40.