Mathematical Modeling and Algorithm Application
ISSN: 3006-0842 | Vol. 4, No. 1, 2025

Research on Fresh Vegetable Market Sales Characteristics and Prediction Model Based on Time Series

Junjie Ma[#], Mohan Pan[#], Youlun Zhang[#]

School of Information Engineering, Jincheng College, Nanjing University of Aeronautics and Astronautics, Nanjing 211156, China.
[#]These authors contributed equally.

Abstract: Due to the short sales time and storage difficulties of fresh agricultural products, how to price and replenish future fresh products based on historical sales has become an important research topic. This article studies the optimization of sales strategies for fresh vegetables based on time series prediction models and data visualization analysis. Through data collection and preprocessing, detailed sales data from multiple vegetable markets across the country were obtained. Data visualization analysis shows that flowering and leafy vegetables account for about 42% of total vegetable sales, followed by chili plants at 19% and edible fungi at 16%. Taking Hong hu lotus root as an example and using the LSTM time prediction model, it is forecasted that when the supply time is 800 days, the maximum replenishment quantity of crispy lotus roots is about 120,000; when the supply time is 1,000 days, the sales volume will reach the second peak of 100,000. These visualizations provide support for developing targeted sales strategies.

Keywords: Data visualization analysis, Cluster analysis, LSTM time prediction, Time step pulse.

1. Introduction

In modern society, with the continuous improvement of people's living standards, consumers' requirements for food quality are also increasing, especially in fresh vegetables. Fresh vegetables, as an important component of daily diet, pose significant challenges to the supply chain management of fresh supermarkets due to their short shelf life and susceptibility to spoilage. Therefore, fresh food supermarkets must effectively manage their replenishment and pricing strategies to meet consumer needs while maximizing profits. Firstly, to solve the problem of analyzing the distribution patterns and interrelationships of sales volume of various vegetable categories and individual products, an evaluation index system for sales volume of various vegetable categories and individual products is established. Multiple analysis methods are used to explore the correlation between individual products, and finally, the correlation between vegetable categories and varieties is obtained through hierarchical clustering analysis. Next, to solve the problem of analyzing the relationship between the total sales volume and cost markup pricing of various vegetable categories and predicting the total replenishment volume and pricing strategy for the next week, regression analysis is used to determine the linear relationship between the total sales volume and cost pricing and a time series prediction model is established to predict the total replenishment volume and pricing strategy for the next week. Finally, regarding the further development of single-item pricing, while meeting the market demand for various vegetable categories, the grey prediction model is used to optimize the model by adding constraint conditions. The grey prediction model is used to predict the total replenishment amount and pricing strategy for a future day, to maximize the revenue of supermarkets, and the penalty function is used to improve the accuracy of the prediction model.

2. Overview of Vegetable Variety Evaluation and Sales Analysis

2.1. Data preprocessing

The research data in this article is sourced from www.mcm.edu.cn. Based on the raw data, this study first needs to clarify the distribution pattern of sales among various categories and individual vegetable products. To discover this pattern, this paper uses Matlab for data processing, adding up the sales of the required categories and individual products within three years.

The following summation formula was used:

$$sum = \sum_{i=1}^{n} (a_n) \tag{1}$$

Among them, sum is the total sales amount of each item or category, n is the amount of data required, and an is the sales amount corresponding to 1 to n.

Finally, use Excel to convert the requested sales data into pie charts and scatter plots. The distribution pattern can be intuitively represented through visual analysis [1] The following figure:

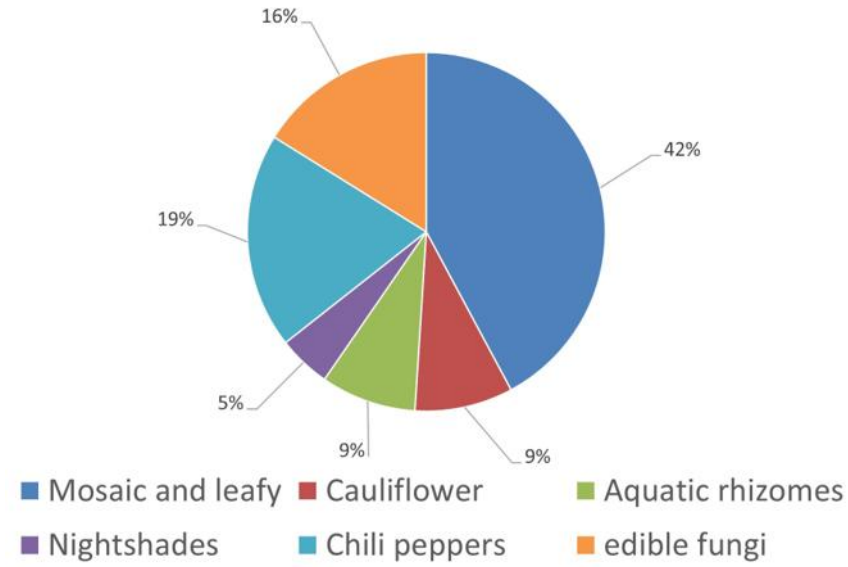


Figure 1: Distribution of Total Sales of Similar Products

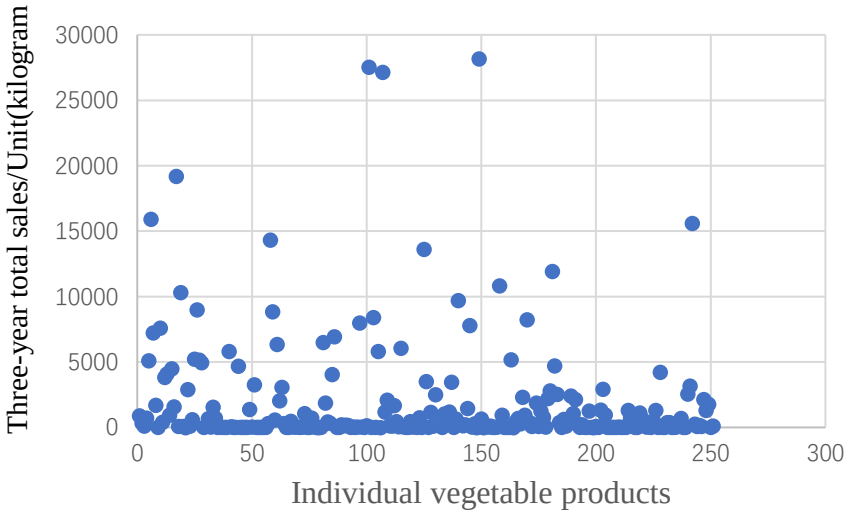


Figure 2: Distribution of Total Single Product Sales

According to Figure 1, it can be observed that flowering plants account for the largest proportion of 42% among the six categories, while eggplants account for the smallest proportion of 5%. Edible fungi and chili peppers account for approximately equal proportions of 16% and 19%, respectively. From this, it can be analyzed that flowering plants are the most popular among people, while eggplants are relatively less common.

According to Figure 2, it can be observed that the horizontal axis represents 251 individual vegetable products, and the vertical axis represents the total sales over the past three years. The difference in sales volume of individual products is not significant, with most sales concentrated below 5000 kilogram and a very small number exceeding 10000 kilogram.

2.2. Cluster analysis model

This article uses hierarchical clustering [2] to cluster various categories, individual products, and sales volume of vegetables as indicators. Firstly, unify variables and perform data processing. Through clustering, clustering groups are obtained, and those in the same group have a strong correlation. Obtain the genealogy diagram as shown in Figure 3.

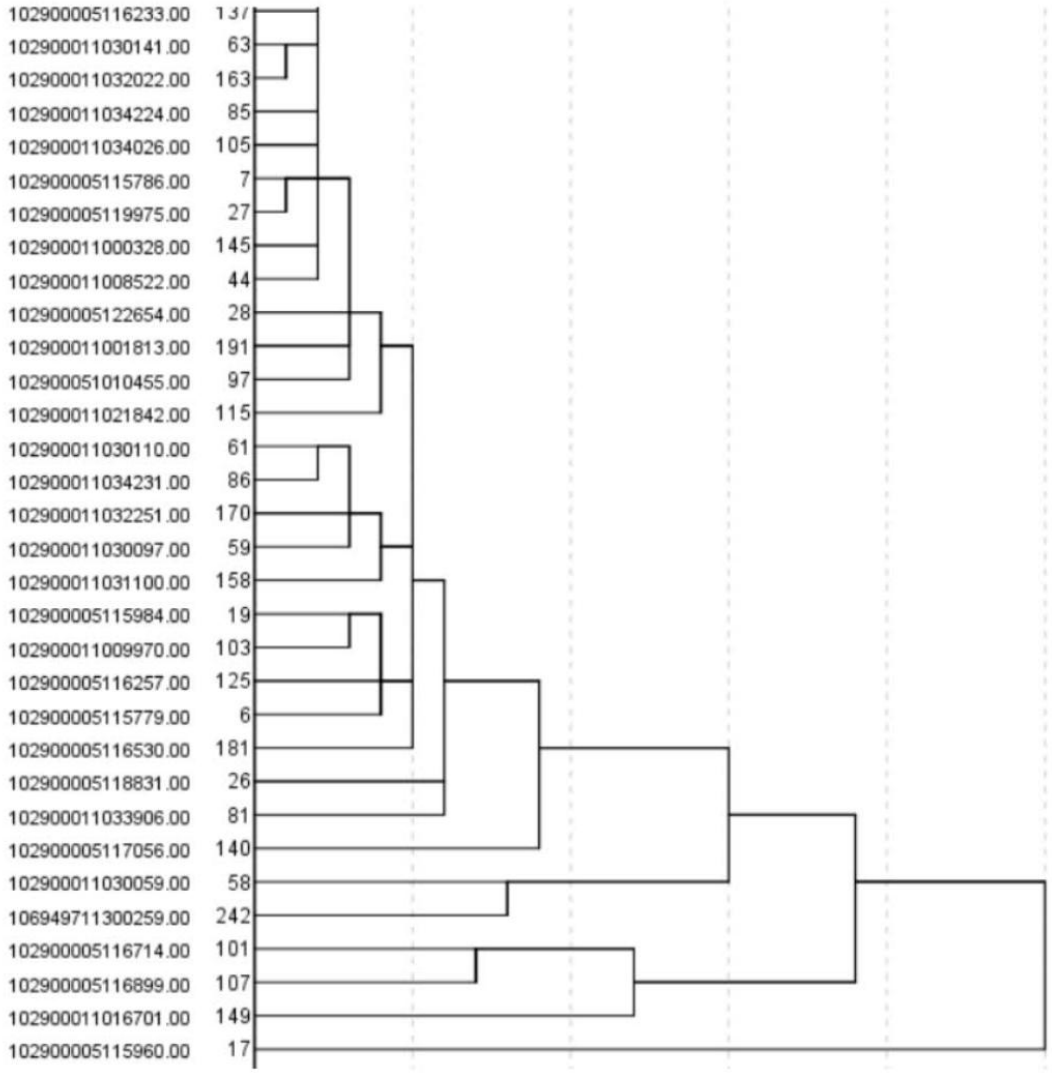


Figure 3: Genealogy of each product

Based on comprehensive analysis of multidimensional data such as sales volume, market penetration rate, and consumer preferences of vegetable categories, we have accurately divided them into three major segmented markets: firstly, high-frequency daily consumption vegetable categories, which have a wide consumer base, affordable price positioning, and high cooking adaptability, such as representative varieties such as Chinese cabbage, showing a huge sales volume and stable market position in the market; Next is the auxiliary vegetable category, represented by broccoli, lotus root, Wuhu green pepper, etc. Although they do not dominate the dishes, they play an indispensable role in enriching the flavor and enhancing the overall dining experience with their unique taste and nutritional value, demonstrating strong potential for dish pairing; Finally, there are nonmainstream or specialty vegetable categories, which have significant differences in sales and usage frequency compared to the first two categories due to their unique taste characteristics, relatively high price positioning, or specific cooking skills requirements. However, they still have a certain market appeal and may demonstrate market segmentation potential for specific consumer groups or catering scenarios. Through this refined market classification, enterprises can more accurately capture market dynamics and consumer preferences, and then develop and implement effective market strategies and product portfolio optimization plans.

3. Using LSTM for predicting future retail sales

3.1. Establishment and Solution of Time Prediction Model

On a category-by-category basis, By using the LSTM time series prediction model [3], analyze the relationship between sales volume and pricing of product categories, and predict the sales volume of supermarkets in the next seven days. To achieve this goal, this article uses code to standardize and sample the data.

The formula for forgetting the gate is as follows:

$$f_t = \sigma(W_f[x_t, h_{t-1}] + b_f) \tag{2}$$

Among them, W_f is the weight matrix of the forget gate, b_f the bias term, x_t is the input of the current time step, h_{t-1} the hidden state of the previous time step, σ the sigmoid function, and maps the input value to a probability value between 0 and 1.

The input gate formula is as follows:

$$i_t = \sigma(W_i[x_t, h_{t-1}] + b_i) \tag{3}$$

Among them, W_i is the weight matrix of the input gate, b_i which is the bias term.

The output gate formula is as follows:

$$o_t = \sigma(W_o[x_t, h_{t-1}] + b_0) \quad (4)$$

Among them, W_o is the weight matrix of the output gate, b_0 is the bias term.

Firstly, load the data. This article analyzes the 120th item, with the horizontal axis representing the number of days and the vertical axis representing the total replenishment amount. The daily replenishment amount of Hong hu lotus root (crispy lotus root) for the 120th item is obtained, as shown in Figure 4:

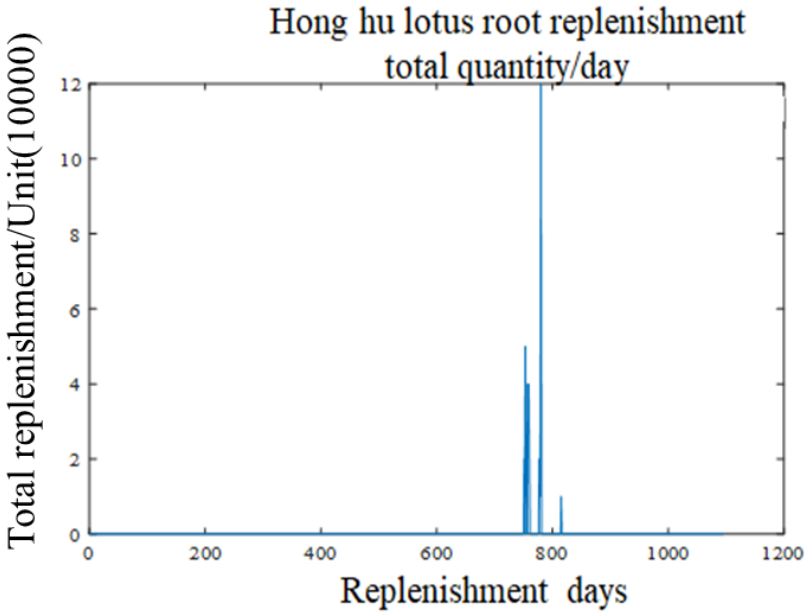


Figure 4: Replenishment of Crispy Lotus Roots within Three Years

From Figure 4, it can be seen that the daily total yield of Hong hu lotus root did not show a significant peak or a significant downward trend, indicating that its yield maintained a relatively stable growth trend in the time series. Specifically, over time, the yield shows a gradually accumulating and slowly increasing trend, which may be due to the long growth cycle of Hong hu lotus root and the gradual increase in yield influenced by various environmental factors. From this chart, it can be observed that the total supply quantity can reach 120000 units when the supply date is around 800 days.

In predicting future time steps, use the predicted values to plot the training time series [4] as shown in Figure 5:

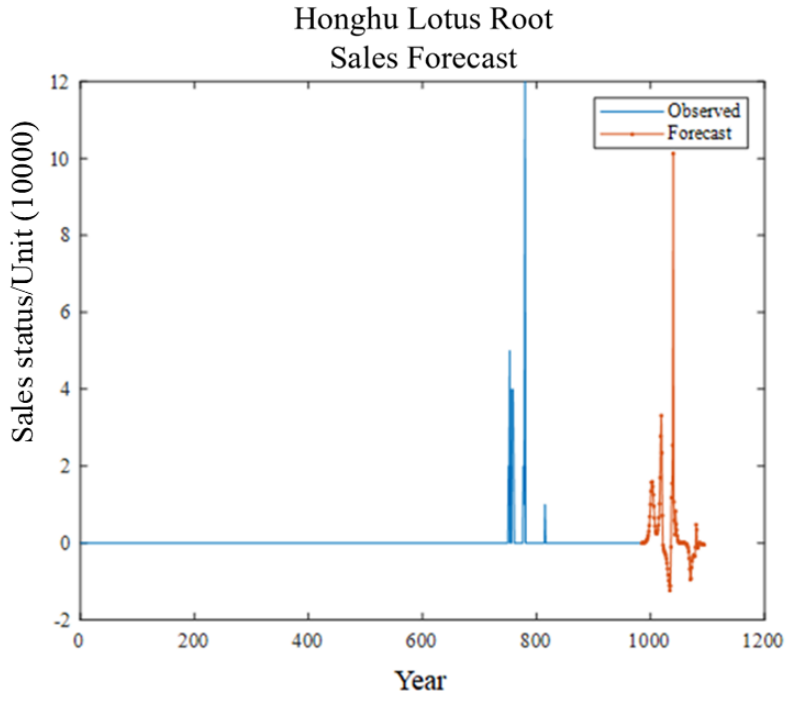


Figure 5: Prediction of Future Sales of Crispy Lotus Root

From Figure 5, it can be seen that sales will reach the second peak period of 100000 units when the supply time is 1000 days.

In predicting the future time step[5], the comparison between the predicted value and the test data is shown in Figure 6, and it can be concluded that the root mean square error is equal to 1.2487.

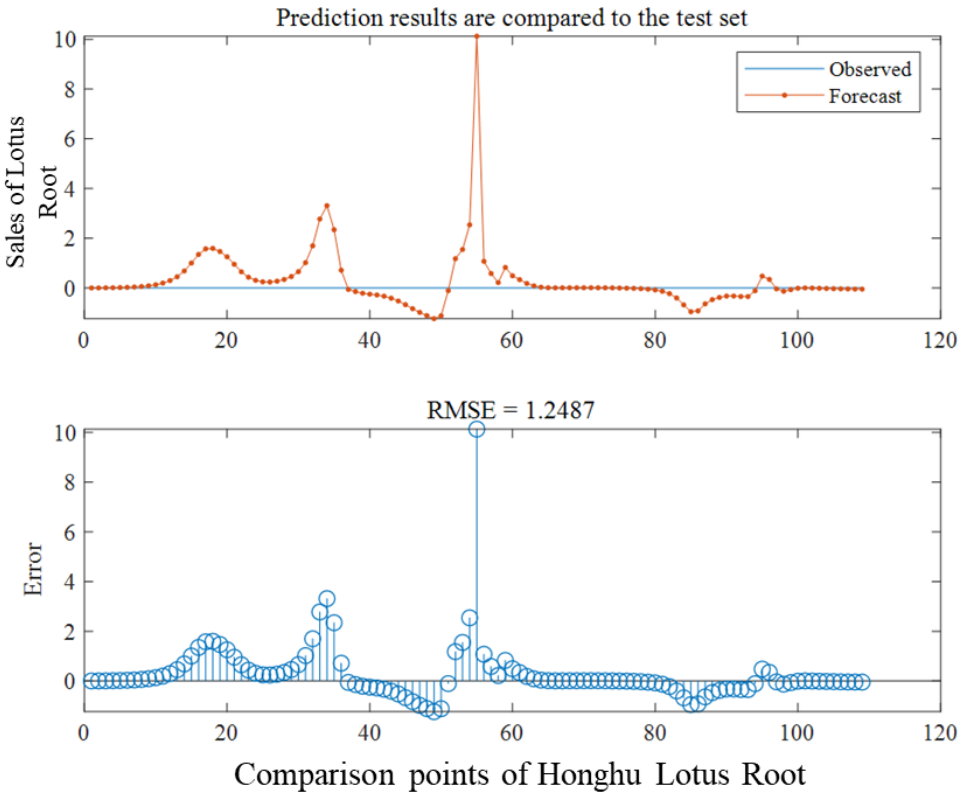


Figure 6: Comparison of predicted values and test data for crispy lotus roots

Figure 6 shows the comparison between the predicted results and the test set, as well as the comparison of specific points. The difference between predicted sales and actual sales is shown through a red line chart. The horizontal axis of the line chart represents time or series number, and the vertical axis represents sales revenue. By observing the chart, it can be found that there is a certain deviation between the predicted value and the actual value, which reflects the accuracy of the prediction model in predicting sales revenue. The distribution of prediction error within a specific range is shown through a blue scatter plot, resulting in a root mean square error value of 1.2487.

As shown in Figure 7, the root mean square error obtained by updating the network state with observed values is 0.0023.

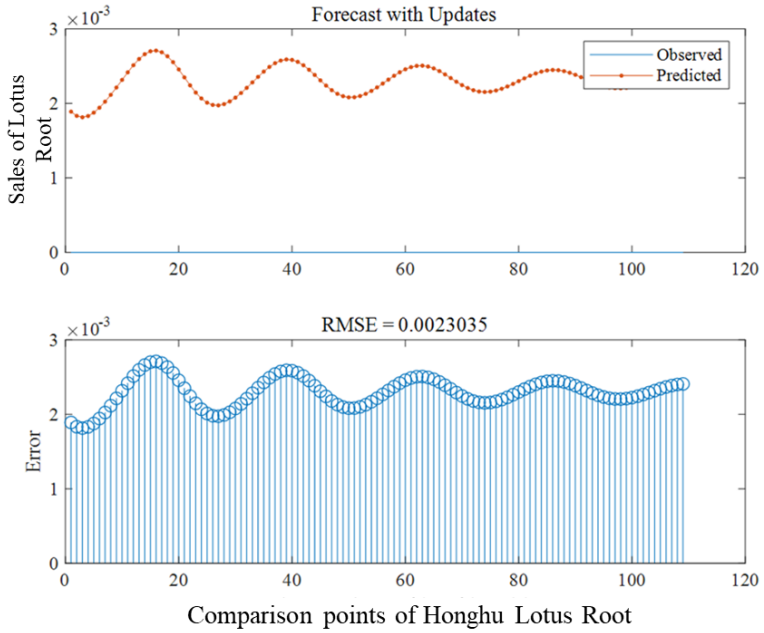


Figure 7: Crispy Lotus Root Sales Prediction Network

Figure 7 clearly shows the differences and trends between actual sales data and predicted sales data through a time series comparison chart, and provides the evaluation indicator RMSE for prediction accuracy. Predict the data for the next 7 days, as shown in Figure 8:

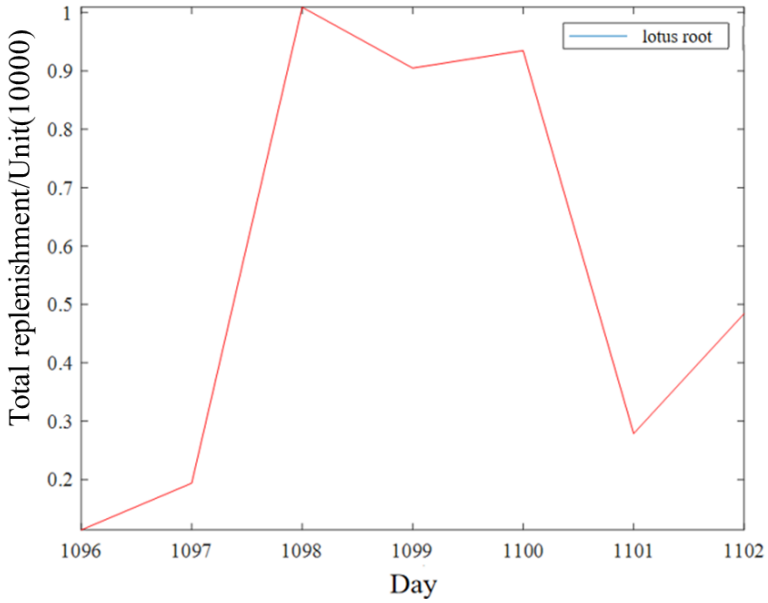


Figure 8: Seven-day replenishment of Hong hu lotus roots

Figure 8 shows the trend of replenishment volume through a line chart, and specifically points out the supply volume for Hong hu lotus root. From the 1096th to the 1098th day, the supply of Hong Hu lotus roots significantly increased, from replenishing 2000 Hong Hu lotus roots per day to replenishing 10000 Hong Hu lotus roots per day. From the 1099th to the 1101th day, the replenishment of lotus roots showed a fluctuating downward trend, with the daily replenishment decreasing from 9000 Hong Hu lotus roots per day to 3000 Hong Hu lotus roots per day. Although there have been significant fluctuations in the replenishment of lotus roots over the past seven days, the overall trend is showing an upward trend.

4. Conclusions

This study explores fresh vegetable sales strategy optimization based on time series prediction models and data visualization analysis. Through multi-dimensional analysis, we reveal the sales patterns, trend changes, and external correlations across vegetable categories. Data shows leafy vegetables dominating the market share (42%), followed by chili peppers (19%) and edible fungi (16%). Using Honghu lotus root as a case study, we applied the LSTM time series prediction model for sales forecasting. Results show replenishment volume peaking at an 800-day supply cycle (approximately 120,000 units), with sales volume reaching a second peak at 1,000 days (approximately 100,000 units). The LSTM model demonstrates outstanding performance in sales forecasting, precisely capturing temporal dependencies and nonlinear characteristics in sales data. Through validation, the model achieved an RMSE of 1.2487 in future predictions, decreasing to 0.0023 after calibration, confirming its high accuracy. The study's innovation lies in developing a multi-dimensional visualization framework, systematically presenting sales characteristics, trends, and external factor interactions through various charts, providing theoretical support for precise sales strategy development.

While this study has achieved certain results, it is limited by the narrow geographical coverage of data samples, and the existing replenishment and pricing strategies overly focus on overall trends and profit maximization while neglecting consumers' differentiated demands. Future research could enhance the model's practicality and prediction accuracy by expanding data collection scope, establishing dynamic update mechanisms, and deepening personalized demand analysis, thereby providing more reliable reference for fresh retail enterprises' business decisions.

References

[1] Li Deng feng, Yang Shuang Huan, Wang Xiaohong, etc Research on Large scale Heat Network Scheduling Strategy Based on Visualization Technology of Heating Effect[J]. Regional heating, 2024, (06): 8-13+25.

[2] Li Yin Na, Yun Ke Li, Zhang Yu, etc Cluster analysis based on the comparison of toxic, warm yang, medicinal properties, taste, meridian and efficacy[J]. Chinese folk therapy, 2024, 32(24): 7-12.

[3] Zhang Wen, Mao Wen Tao, Wang Yu, etc Periodic time series prediction based on robust information fusion LSTM[C]// Northeastern University, Information Physics System Control and Decision making Professional Committee of the Chinese Society of Automation. Proceedings of the 35th China Conference on Control and Decision making, 2023: 6.

[4] Sun Jin Xin, Zhong Jian, Yao Lei, etc Highway traffic flow prediction based on EEMD Att GCN LSTM[J]. Industrial control computer, 2024, 37(12): 39-41.

[5] Li Qian peng, Jia Shun Cheng, Zhang Tie Lin, etc Efficient image classification based on adaptive time step pulse neural network [J] Journal of Automation, 2024, 50(09): 1724-1735.