A Fresh Vegetable Optimal Replenishment Model Based on LSTM-ARIMA

Qianrong Zhang[1, 2], Hazirah Bee[1,*]

[1] Faculty of Information Technology, City University Malaysia, Darul Ehsan, 46800, Malaysia
[2] College of Big Data and Artificial Intelligence, Chengdu Technological University, Chengdu, 610000, China
* Correspondence: Hazirah Bee

Abstract: This study examines how supermarkets can adjust future pricing decisions and replenishment quantities based on historical dish sales and sample sales data. Then, a model is established based on factors such as cost-plus pricing and profit maximization to determine the optimal pricing strategy and replenishment quantity. Firstly, the top 20 most popular items are selected and ranked. Next, the annual sales of these 20 single items corresponding to each type of dish are presented in a histogram, and their sales trends and distribution patterns are analyzed. Then, Q-Q plots for the sales of each type of dish and each single item are conducted to perform normality tests. It is found that the dishes follow a normal distribution, while the sales volume of individual items does not. Pearson and Spearman correlation coefficients are used for correlation analysis, revealing a certain degree of correlation between edible fungi and aquatic root and stem types, with a correlation coefficient of 0.65; the correlation between sweet potato shoots and bamboo leaf vegetables is higher, at 0.91. We first convert each indicator value of each category based on the relationship between total sales and cost-plus pricing indicators. We establish a multivariate regression equation based on the least squares method from aspects such as purchase price, selling price, addition coefficient, and loss rate to the total sales volume, and standardize the values to solve the weight coefficients in each factor's regression equation, obtaining the final multivariate regression equation. Then, an LSTM-ARIMA model is established, using the first 90% of the samples as the training set for fitting and parameter adjustment, obtaining an ARIMA (2, 0, 2) model, and predicting the daily replenishment quantity from July 1st to 7th, obtaining the final optimal pricing result, and calculating the maximum profit to be 4664.493 yuan.

Keywords: replenishment quantity; pricing strategy; correlation analysis; lstm-arima; maximum profit

1. Introduction

The insurance period for vegetable products is relatively short, and their appearance quality gradually deteriorates over time. Therefore, if most varieties of vegetables are not sold on the same day, they cannot be sold the next day. Thus, for supermarkets selling vegetables, they need to replenish stocks daily based on the historical conditions of different products. The vegetables sold by supermarkets come in various varieties and have different origins. Due to their purchase and sales time usually being from 3 to 4 a.m., the merchants need to make replenishment decisions without knowing the specific individual items and purchase prices for the day. It is now known that the pricing of vegetables is generally done using the "cost-plus pricing" method; supermarkets will offer discounts on vegetables with poor appearance; the sales volume of vegetables is in a parallel relationship with time; the supply varieties of vegetables are more abundant from April to October; and the selling space of supermarkets has certain limitations. Therefore, formulating an optimal replenishment strategy is crucial for the operation of supermarkets [1-3].

In recent years, scholars at home and abroad have conducted a large number of studies on the inventory and replenishment issues of fresh agricultural products, especially vegetable products. Due to the characteristics of vegetables, such as short shelf life, large fluctuations in demand, and high loss rate, although traditional deterministic inventory models (such as EOQ model, newsboy model) can describe the order-inventory relationship to a certain extent, they show obvious limitations when facing frequent fluctuating demand and dynamic price changes [4,5]. Therefore, the research has gradually shifted from static inventory control to demand-driven dynamic replenishment decisions. Some scholars use time series methods (such as ARIMA [6], exponential smoothing model) to approximate model the sales volume of fresh produce, using historical sales data to depict seasonality and trend; there are also studies that combine external factors such as price, promotion, and weather to construct multiple regression models to analyze the elasticity relationship between price and sales volume. These methods have advantages in terms of clear structure and strong interpretability, but their ability to depict nonlinear fluctuations and sudden changes is relatively limited.

With the development of machine learning methods, more and more studies have introduced neural network models into the field of fresh sales prediction and intelligent replenishment [7]. Among them, recurrent neural networks such as LSTM are widely used for sales volume prediction of high-volatile products due to their ability to remember time-dependent relationships; at the same time, "hybrid prediction models" that combine statistical models and deep learning models have gradually become a hot topic, such as the combination of LSTM and ARIMA, to balance the linear structure characterization ability and nonlinear fitting ability. In replenishment decisions, some studies further embed the prediction results into profit maximization or cost minimization frameworks, combining loss rate, cost-plus pricing, and inventory constraints, to form a prediction-pricing-replenishment integrated decision-making system. However, existing studies mostly focus on a single category or ignore the correlation between categories, and there is still less joint analysis of multi-category vegetables in actual supermarket scenarios [8,9]. Based on this, this paper constructs an LSTM-ARIMA prediction model based on correlation analysis and combines it with cost-plus pricing and profit models to form an optimal replenishment and pricing strategy for multi-category vegetables.

Therefore, this paper first analyzes the relationships between different categories and individual items and the distribution patterns of various vegetable categories, and since there are too many single varieties, they can be screened based on whether they are valuable for correlation analysis. Then, it analyzes the distribution patterns and interrelationships of major category indicators and the distribution patterns and interrelationships of individual sales volumes. Since sales volume is a continuous variable and by testing whether the normal distribution of each category item satisfies, then use the spearman or pearson method for correlation analysis, if the vegetable category is a discrete variable, the chi-square test can be used for analysis, and the sales volume of individual items can be related to consumers' subjective combinations and analyzed using clustering. In addition, this article analyzes each major category as a unit, without considering individual items. It does not consider the relationship between sales volume, cost, and pricing. Instead, it separately analyzes each major category, summarizes the results, and examines the relationship between the daily sales volume changes and pricing. An LSTM-ARIMA model is established, and based on the previous data, it predicts the required replenishment quantity for each category in the future. At the same time, the maximum profit under this replenishment strategy is calculated.

2. Methodology

2.1 Model for Correlation between Dishes and Sales Volume

2.1.1 Pearson correlation model

The Pearson correlation coefficient is a measure of linear correlation. We define the Pearson correlation coefficient as follows $\lambda_{X,Y}$: When the data show a non-linear relationship or do not follow a normal distribution, this correlation analysis method can be used for calculation, which can be expressed by the following formula [10].

$$\lambda_{X,Y}=\frac{E(XY)-E(X)E(Y)}{\sqrt{E(X^2)-E^2(X)}\sqrt{E(Y^2)-E^2(Y)}} \tag{1}$$

Among them, the covariance is represented cov, and the standard deviations are indicated δ. The calculation formula for the expected value is as follows:

$$E(X)=\frac{1}{n}\sum_{i=1}^{n}X_i \tag{2}$$

The solution of variance is as follows:

$$\delta_X^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2 \tag{3}$$

The solution of covariance is as follows:

$$\mathrm{cov}(X,Y)=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y}) \tag{4}$$

The strength of the Pearson correlation coefficient can be represented by α, and the calculation of its strength follows the following formula:

$$\alpha=\frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}} \tag{5}$$

Meanwhile, the standards for the strength of correlation are shown in Table 1.

Table 1. Standards for Correlation Strength

$\lvert\alpha\rvert$	**Correlation Strength**	$\lvert\alpha\rvert$	**Correlation Strength**
0.8-1.0	Highly correlated	0.2-0.4	Weak correlation
0.6-0.8	Strong correlation	0.0-0.2	Uncorrelated
0.4-0.6	Correlation		

2.1.2 Establishment of the spearman correlation model

The Spearman correlation coefficient is a non-parametric measure of rank correlation. We define the Spearman correlation coefficient as d_t. When the data exhibit a non-linear relationship or do not follow a normal distribution, this correlation analysis method can be used for calculation.

2.2 Sales Volume Prediction Model Based on LSATM-ARIMA

The ARIMA model is used to model and predict the time series data. Then, the LSTM model selects the first 90% of the samples as the training set for training, and the remaining 10% of the samples are used as the simulation values for prediction simulation. The simulated values are compared with the real values, and the relevant parameters are adjusted. Then, the relevant parameters are used in the ARIMA model to predict the total sales volume.

2.2.1 Establishment of the LSTM Model

Its working principle mainly involves introducing gating mechanisms to solve the problem that RNN cannot handle long-term dependencies. To avoid the situation where information experiences gradient disappearance and cannot have long-term dependencies, three gating mechanisms, namely the input gate, the forget gate, and the output gate, are introduced. These three gates can effectively control the flow and forgetting of information, thus solving the problem of gradient disappearance and enabling long-term predictions. The specific workflow is shown in Figure 1 below:

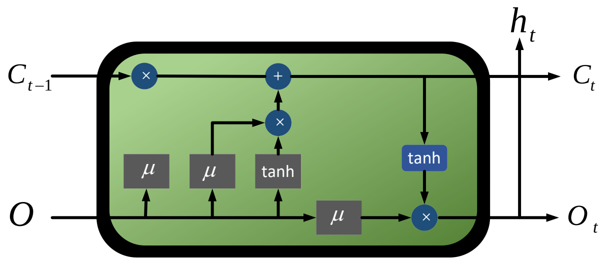

Figure 1. LSTM Workflow

Each gate is characterized by: consisting of a sigmoid neural network layer and pointwise multiplication operation. The sigmoid layer outputs numbers between 0 and 1, describing the amount each component should pass through [11]. When it is 0, it indicates "do not allow information to pass through", and when the value is 1, it indicates "allow all information to pass through".

Step 1: The Forget Gate: Mainly controls the forgetting of certain elements in the old state, and is represented by the following formula:

$$\hat{f}_t = \mu(V_f[x_t, o_{t-1}] + a_f) \tag{6}$$

Among them, $\hat{f}_t$ represents the sigmoid function, which is used to determine which elements in the control state are to be forgotten and selected. a_f represents the bias term, x_t represents the input of the current time state, o_{t-1} represents the hidden state in the previous time, and μ represents the probability value determined by the sigmoid function for the forgetting selection of state elements, that is $\mu \in [0,1]$.

$S_t \to 1$ indicates that past state information should be retained more; $S_t \to 0$ indicates that past state information should be forgotten more.

Step 2: Input gate: Controls the flow of new input information. At each time step, the LSTM model receives a new input value and calculates its new candidate state, which is the new information. The flow of this information over time steps is controlled by the input gate. It is represented by the following equation:

$$\hat{i}_t = \mu(V_i[x_t, o_{t-1}] + a_i) \tag{7}$$

Among them, V_i represents the weight matrix of the inputs, b_i is the bias term, x_t represents the input at the current time state, o_{t-1} represents the hidden state at the previous time, μ represents the probability value determined by the sigmoid function for deciding which state elements to forget, that is $\mu \in [0,1]$.

The LSTM model calculates the state of the candidate information. We denote it as $\hat{c}_t$, which represents the meaning as: How much influence can the new input information at the current time step have on the candidate state? The equation describing this information state is as follows:

$$\hat{c}_t = \tanh(V_c[x_t, o_{t-1}] + a_c) \tag{8}$$

Among them, $\tanh$ is the hyperbolic tangent function, which can limit the input value within the range of [-1, 1].

Step 3: Information State: This is the core component of the entire LSTM model network, responsible for information storage and transmission, acting as an information transfer station. It can also control the flow and update of information. The following information processing equations are as follows:

$$C_t = \hat{f}_t * C_{t-1} + \hat{i}_t * \hat{c}_t \tag{9}$$

This equation represents the change in the updated state of the information at each time step. Here, $\hat{f}_t$ represents the forget gate, which is a weight used to measure the degree of information forgetting; $\hat{i}_t$ represents the input gate, $\hat{c}_t$ represents the weight for updating the information state; it represents the information state of the candidate elements at the current time step, it represents the extent to which the new value input will affect the information state.

Step 4: Output gate: Controls the output of the information. The LSTM model needs to output the valid information in this state, and requires the output gate to control and coordinate the output of the information. The following equation represents its working principle:

$$\hat{o}_t = \mu(V_o[x_t, o_{t-1}] + a_o) \tag{10}$$

LSTM will process the previous information state value C_t through the function $\tanh$ to obtain the current time step hidden state f_t, and the processing formula is as follows:

$$f_t = \hat{o}_t \cdot \tanh(C_t) \tag{11}$$

2.2.2 Establishment of the ARIMA Model

The ARIMA model is also known as the Auto regressive Integrated Moving Average model. It generates a data random sequence based on the temporal changes of the prediction target, and establishes a related auto regressive model to approximately simulate this sequence. It is then explained using a mathematical model. After continuously modifying the coefficient values, this model based on the time-varying sequence can be used to predict future values based on historical values. It can be represented by the following formula:

$$[1 - \sum_{i=1}^{q} \theta_i I^i](1 - I)^d X_i = [1 + \sum_{i=1}^{p} \phi_i I^i]\alpha_i \tag{12}$$

Among them, I is the lag operator, α_i is the white noise sequence, q, p, d are the three parameter values in the model, q represents the lag number of the time series itself, d represents the order value of the original sequence after differencing to become a stationary sequence, p represents the lag number of the prediction error. This model can be expressed as ARIMA(q,p,d). At the same time, we use the BIC Bayesian Information to measure the complexity and fitting effect of the model. The expression is as follows:

$$BIC = \ln(i)n - 2\ln(K) \qquad (13)$$

Among them, i represents the sample size, n represents the number of variables in the model, and K represents the maximum likelihood value under this model. The coefficients are adjusted and corrected mainly based on probability through the Bayesian equation, so as to make the model's prediction effect more consistent with the actual value, and at the same time avoiding the influence of subjectivity.

Table 2. Number of vacancies for the top 20 categories

No.	Category	Number	No.	Category	Number
1	Broccoli	0	11	Shiitake Mushroom	3
2	Lotus Root	0	12	Baby Chinese Cabbage	3
3	Purple Eggplant	0	13	Spinach	3
4	Leaf Lettuce	1	14	White Beech Mushroom	3
5	Long Green Chili Pepper	1	15	Romaine Lettuce	4
6	Pak Choi	1	16	Water Spinach	4
7	Baby Pak Choi	1	17	Green Bell Pepper	6
8	Green Eggplant	1	18	Red Bell Pepper	7
9	Chinese Cabbage	2	19	Round Eggplant	8
10	Choy Sum	2	20	Sweet Potato Leaves	9

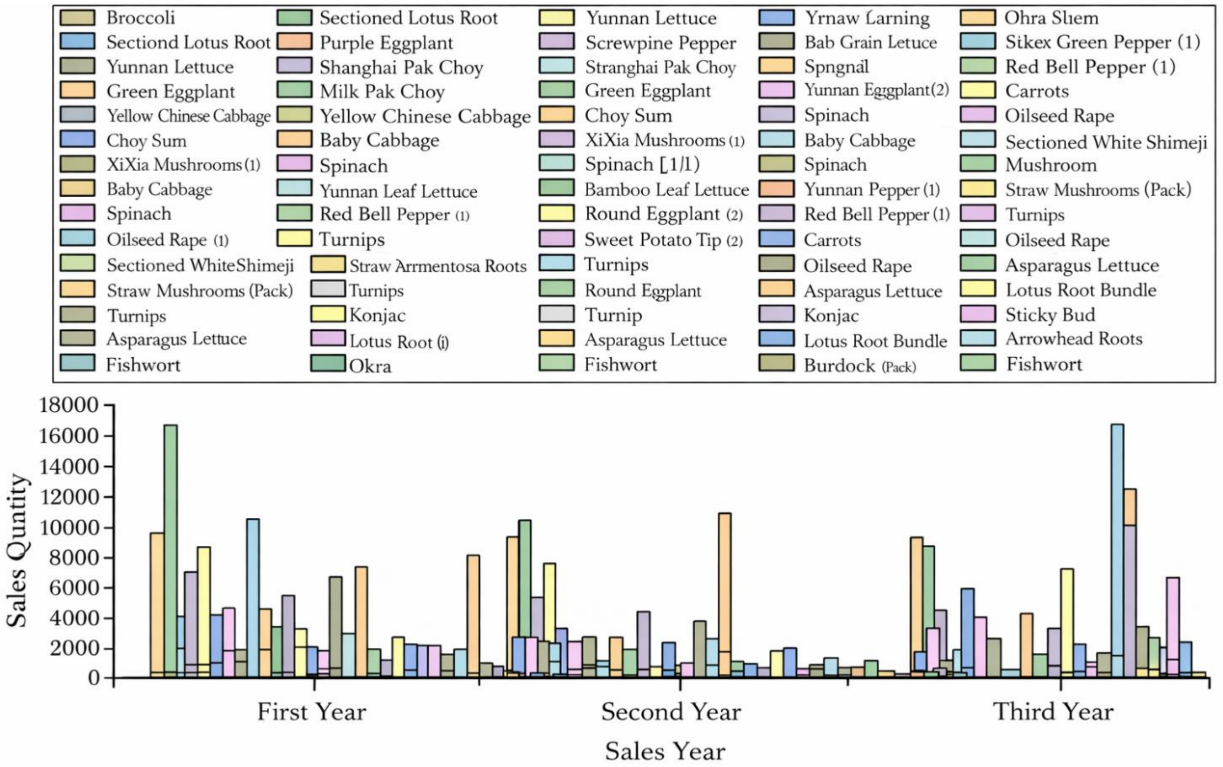

Figure 2. Bar chart showing the annual sales trend of individual products

3. Data Collection and Analysis

First, consider the distribution patterns of sales volumes for each category and individual items of vegetables. Then, conduct a correlation analysis among them. Considering that there are too many types of individual items, this causes certain difficulties for the subsequent analysis and solution of the problem. We analyze the sales of each type of individual item in each month from 2020 to 2023, and sort the items with the least number of sales vacancies per month. The top 20 items are selected for further analysis as samples. We believe that analyzing the relationship between sales and dishes is more effective. The corresponding vacancy times are shown in Table 2.

The 20 selected items were analyzed in terms of their distribution patterns among various vegetable categories. The total monthly sales volume of each category of items from 2020 to 2023 was statistically recorded on an annual basis, and a bar chart was created to display the trend of annual sales volume of the items, as shown in Figure 2:

As can be seen from Figure 1, the sales volume of Chinese cabbage was the highest in the first year, the sales volume of Wuhu green peppers (1) was the highest in the second year, and the sales volume of shiitake mushrooms (boxes) was the highest in the third year.

4. Result

4.1 Correlation between Dishes and Sales Volume

We took the time "month" as the continuous variable and conducted correlation analyses between the sales volumes of various vegetable categories and the sales volumes of each individual item. We first performed normality tests and then selected appropriate correlation models for solution. We used QQ plots to conduct normality tests for the major categories and variety indicators, and obtained the following results in Figures 3 and 4. From the Figure 3, it can be seen that the line segment is closer to the point. The rQ values are all close to 1, and the major category indicators follow a normal distribution. Therefore, the Pearson correlation coefficient method can be used to solve for the correlation.

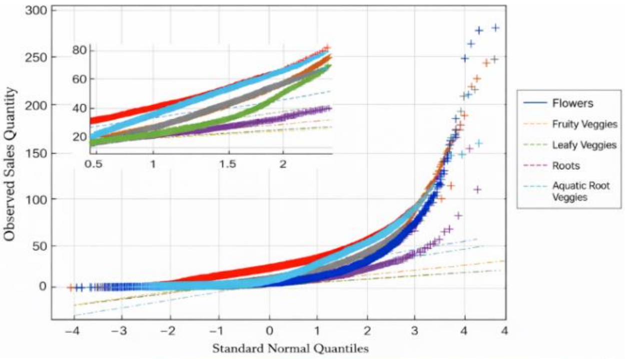

Figure 3. Large-category indicator Q_Q diagram

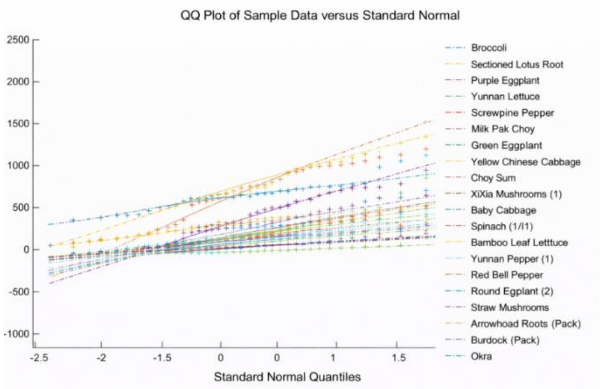

Figure 4. Single product qq chart

From Figure 3, it can be analyzed that the individual item indicators do not follow a normal distribution. Therefore, the Spearman correlation coefficient method can be used to solve for the correlation. This paper calculates the Spearman correlation coefficient between the sales volumes of vegetable dishes and the Spearman correlation coefficient between the sales volumes of each type of individual item, and obtains the correlation coefficient graphs as shown in Figure 5 below:

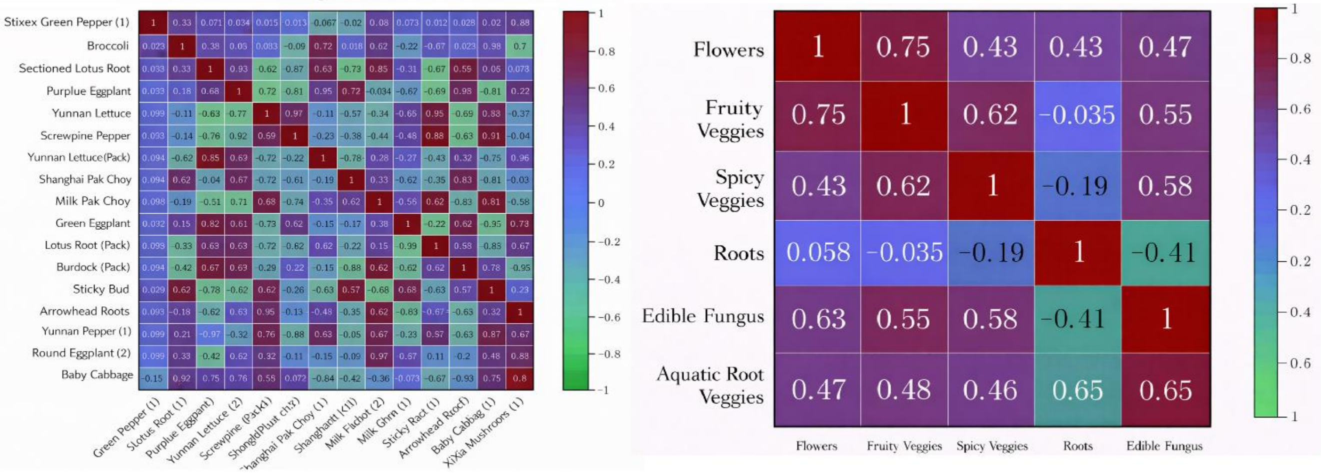

Figure 5. Correlation coefficient heatmap

4.2 Performance Results of LSTM-ARIMA Model

This paper uses the sales data of various vegetables for model training, and uses 90% of the sample data as the training set to simulate LSTM. The simulation situation is shown in Figures 6 below:

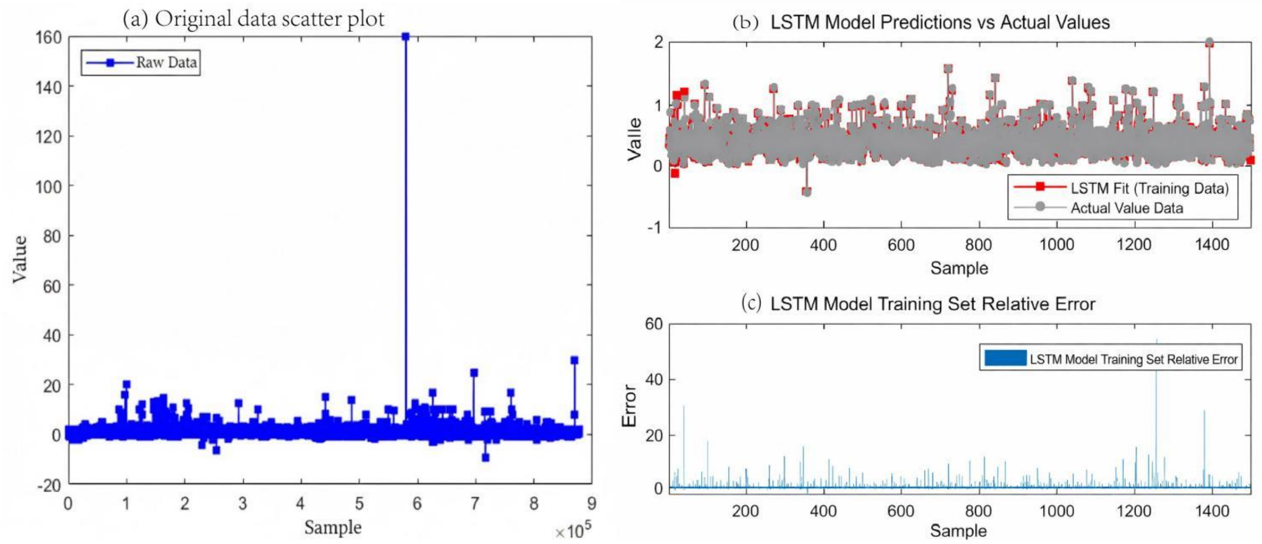

Figure 6. LSTM training set prediction results and errors

The predicted values of the post-prediction model are in excellent agreement with the initial values. The relative errors between the training set and the test set are very small, approximately 0.005. The prediction results are very good.

The dataset of total dish sales required for the Arima model training set, and MATLAB is used to implement the BIC function to price the model. Finally, the solution is:the value reached by BIC was the smallest: -1.919. Finally, the ARIMA model is determined as: ARIMA (2, 0, 2). And the sales volume for the next seven days is predicted. The predicted original data and the iteration graph are shown in Figure 7. It can be seen that the corresponding error is small, and the prediction result is relatively accurate.

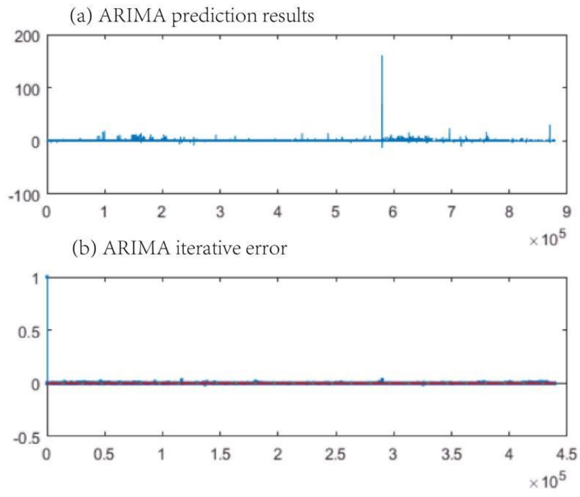

Figure 7. ARIMA (2, 0, 2) prediction results

4.3 Sales Volume and Pricing Forecast

The LSTM-ARIMA model was used to predict the replenishment volume for the next seven days. The forecast results are shown in Table 3. At the same time, this model was also used to predict the pricing for the next seven days. The prediction results are shown in Table 4:

Table 3. Forecast results of various categories' replenishment quantities for the next seven days

Date	Flower-leaf	Flower vegetable	Submerged stem	Brinjal	Chili	Edible fungus
1	83.90	48.560	38.772	37.989	108.292	49.713
2	83.909	48.517	38.743	37.965	108.310	49.659
3	83.919	48.474	38.713	37.942	108.329	49.604
4	83.929	48.430	38.684	37.918	108.347	49.549
5	83.939	48.387	38.654	37.894	108.366	49.494
6	83.949	48.344	38.625	37.870	108.384	49.440
7	83.958	48.301	38.595	37.846	108.403	49.385

Unit: Kilogram

When calculating the final returns of the major categories, the average cost of each major category needs to be calculated. Taking into account the loss rate of vegetables, the average cost of each major category is obtained through cost addition, as shown in Table 5:

The above replenishment quantity, pricing forecast, and average cost have been comprehensively calculated to obtain the final profit: Based on the above formula and the predicted data, the optimal profit is approximately 4664.493 yuan.

Table 4. Prediction results of pricing for major categories in the next seven days

Date	Flower-leaf	Flower vegetable	Submerged stem	Brinjal	Chili	Edible fungus
1	8.269	9.835	9.835	10.435	8.218	9.831
2	8.267	9.835	9.835	10.421	8.211	9.831
3	8.264	9.820	9.835	10.428	8.218	9.832
4	8.262	9.836	9.835	10.422	8.221	9.833
5	8.259	9.947	9.835	10.428	8.218	9.834
6	8.256	9.835	9.835	10.422	8.218	9.835
7	8.254	9.839	9.835	10.428	8.218	9.836

Unit: Yuan /Kilogram

Table 5. Costs of each category

Type	Price
Flower-lea	5.419
Flower vegetable	8.835
Submerged stem	9.899
Brinjal	6.824
Chili	6.686
Edible fungus	8.243

Compared with the average profit of the previous two weeks, it has increased by 350.3 yuan.

5.Conclusion

This paper addresses the issues of uncertain demand, easy spoilage, and reliance on experience in the sales of vegetables in supermarkets, and constructs an integrated model system of "correlation analysis - sales forecast - pricing and replenishment optimization". Through the screening and distribution test of sales data from 2020 to 2023, it is found that the sales volume of vegetable categories approximately follows a normal distribution, while the single-item level shows non-normal characteristics. Based on this, the Pearson and Spearman methods are used to depict the correlation respectively. The results show that the correlation coefficient between edible fungi and aquatic root and stem vegetables is 0.65, and the correlation coefficient between sweet potato tips and bamboo leaf vegetables in the single-item level is 0.91. This verifies that there is a significant linkage relationship between categories, providing a basis for modeling at the category level. In the prediction stage, an LSTM-ARIMA combination model is constructed and the ARIMA(2,0,2) structure is determined. The relative error between training and testing is approximately 0.005. The prediction curve is highly consistent with the real data, indicating that the model has good fitting and generalization ability for the high-volatile time series of vegetable sales. On this basis, combined with cost-plus pricing and loss factors for optimization calculation, the replenishment and pricing strategies for the next 7 days are obtained, and the optimal profit is approximately 4664.493 yuan. In summary, this model not only improves the profit level and reduces the prediction error quantitatively, but also reveals the structural correlation and temporal regularity of vegetable sales qualitatively, achieving a complete closed loop from data analysis to optimization of business decision-making, and has practical application value for the refined management of supermarket fresh produce.

Acknowledgment

This research was supported by Sichuan Provincial Key Laboratory of Philosophy and Social Sciences (CXCYRH25C-38).

Reference

[1] Chen, J., & Dan, B. (2009). Fresh agricultural product supply chain coordination under physical loss controlling. Systems Engineering—Theory & Practice, 29, 54–62.

[2] Emanuele, F., Andrea, M., & Giuseppe, N. D. (2020). Vanilla-option-pricing: Pricing and market calibration for options on energy commodities. Software Impacts, 6, Article 100043. doi.org/10.1016/j.simpa.2020.100043

[3] Liu, J., & Liu, B. (2023). Commodity pricing and replenishment decision strategy based on the seasonal ARIMA model. Mathematics, 11(24), 4921. doi.org/10.3390/math11244921

[4] Wei, F., & Guo, X. (2024). A review of demand forecasting methods for fresh agricultural products. School of Economics and Management, Guangxi University of Science and Technology.

[5] He, Y. (2024). Research on replenishment and pricing strategies for fresh commodities in supermarket. Academic Journal of Mathematical Sciences, 5(1), 41–47. doi.org/10.25236/AJMS.2024.050107

[6] Lu, Z., Nie, W., & Chen, L. (2018). Urban rail transit passenger flow prediction based on ARMA model. Henan Science, 36(5), 646–651.

[7] Liu, X., & Li, S. (2024). A study on vegetable pricing replenishment strategy based on seasonal time series. The Frontiers of Society, Science and Technology, 6(4), 97–104. doi.org/10.25236/FSST.2024.060415

[8] Wang, S. (2017). Research on demand forecasting of cold chain logistics for fresh agricultural products (Master's thesis, Xi'an University of Engineering).

[9] Xu, T. (2023). Research on commodity demand forecasting and supplier evaluation of fresh food distributors (Master's thesis, Xi'an University of Technology).

[10] Liu, Y., Li, M., & Pu, Y. (2023). Replenishment and pricing strategies for vegetable commodities based on optimization class models. Academic Journal of Business & Management, 5(26), 164–171. doi.org/10.25236/AJBM.2023.052625

[11] Pu, Y., Huang, Z., Wang, J., & Zhang, Q. (2024). Research on pricing and dynamic replenishment planning strategies for perishable vegetables based on the RF-GWO model. Symmetry, 16(9), 1245. doi.org/10.3390/sym160 91245