A Study on Vegetable Sales Forecasting Based on Seasonal Indices and LSTM Model

Xudong He[1], Xincheng Huang[2], Siyin Ding[3], Weixing Qian[1, *]

[1]School of Electrical and Automation Engineering, Nanjing Normal University, Nanjing, China, 210023

[2]School of Mathematical Sciences, Nanjing Normal University, Nanjing, China, 210023

[3]School of Geography, Nanjing Normal University, Nanjing, China, 210023

*Corresponding author: 61192@njnu.edu.cn

Abstract. Efficient food supply chain management is crucial for the perishable and seasonally affected vegetable sales field due to global population growth and urbanization acceleration. This trend has led to the widespread application of data analytics and machine learning technologies for optimizing pricing strategies and predicting sales trends. This study aims to explore the relationship between the total sales volume of a vegetable category and its cost-plus pricing using the Pearson correlation coefficient analysis method. Additionally, it seeks to establish a support vector regression (SVR) model and an improved long short-term memory network (LSTM) model based on seasonal index prediction. The projected purchase volume of mosaic and leafy vegetables for the next 7 days is 135.47, 135.65, 127.00, 125.26, 83.11, 105.29, and 157.34 kg. Finally, this dissertation formulated a purchase strategy for the upcoming week based on the predicted sales volume. We calculated the 7-day profit margins to be 0.80, 0.61, 0.62, 0.77, 0.63, 0.57, 0.63. This analysis offers a theoretical foundation for enhancing operational efficiency and profit levels in the vegetable sales sector, as well as practical guidance.

Keywords: SVR Model, Seasonal Indices and LSTM Model, Vegetable Sales Forecast, Machine Learning.

1. Introduction:

With the growing global population and the acceleration of urbanization, the efficiency of the food supply chain has garnered significant attention in both research and practice [1]. Particularly in the field of vegetable sales, where products are perishable and subject to seasonal changes [2-3], pricing strategies, and inventory management pose considerable complexity. In this context, where traditional pricing and inventory management methods are often unable to keep up with the quickly shifting demands of the market, it becomes imperative to investigate more sophisticated approaches for forecasting sales trends and enhancing pricing strategies. This pursuit is essential for creating a predictive model for the sales of fresh vegetables, which can then be utilized to inform and optimize supply and demand dynamics within the vegetable market, ultimately fostering a more balanced economic environment [4-6].

To investigate the relationship between vegetable sales volume and cost-plus pricing, this dissertation initially employed the Pearson correlation coefficient method. Proposed by Karl Pearson, a pioneering figure in contemporary statistics, the Pearson correlation coefficient provides a more precise measure of the linear relationship between two variables. For sales forecasting, this dissertation sequentially utilize the Support Vector Regression (SVR) model and the Long Short-Term Memory Network (LSTM) model based on seasonal index forecasting. The SVR model, developed by renowned computer scientist and statistician Vladimir Vapnik, excels in capturing nonlinear relationships and complex data patterns, making it widely applicable in various commodity purchase tasks. Concurrently, the seasonal index prediction model effectively handles data with evident seasonal variations, and its integration with LSTM and SVR models enhances prediction robustness [7].

Initially, our study applies Pearson's method to analyze category vegetables, revealing a statistically significant negative correlation at the top 1% significance level when considering data with one day as the dimension. Attempts to establish a linear regression equation yield poor linearity. Subsequently, the dissertation successfully depict the relationship between total category sales volume and category vegetable cost-plus pricing through an SVR model. Additionally, an improved LSTM model, based on seasonal index forecast, is developed to predict sales volume over the next 7 days, guiding daily replenishment amounts. Finally, leveraging the model predictions, this article formulate a 7-day pricing strategy to optimize operations and maximize profits.

2. Vegetable sales and cost-plus pricing analysis

2.1. Data preprocessing

This dissertation conducted an investigation on a vegetable store, tracking sales records of six different categories of vegetables and recording their relationships between cost, pricing, and sales volume over a period of time. Six different vegetable categories are denoted as No.1 to No.6, and the date difference is computed using the DATEDIF function, with a focus on the data about the No.1 vegetable category. In the MATLAB environment, two numeric matrices, namely data1 and data2, are created to segregate the date number, item number, wholesale price, and sales price into distinct matrices. Subsequently, data1 and data2 undergo MATLAB processing to ascertain the anticipated sales pricing for each vegetable on a given day and to compute the cost-profit margin. Instances lacking sales data have their corresponding day's data cleared. Ultimately, the cost-profit margin, single product wholesale price, and single product selling price for the same date are averaged, while simultaneously summing the sales volume. Concurrently, the date-coded data is adjusted by subtracting 365 to mitigate the influence of different years on sales volume fluctuations.

2.2. The relationship between total sales volume and cost-plus pricing based on MLR and SVR--Taking Category 1 as an example

The pricing of vegetables sold in supermarkets generally adopts the "cost-plus pricing" method. Cost-plus pricing is a pricing strategy that determines the size of the added profit according to the cost-profit rate, and its mathematical formula is as follows:

$$P = c * (1 + r) \tag{1}$$

Where P represents the unit price of the commodity, c represents the total unit cost of the commodity, and r represents the cost-profit rate of the commodity.

In product cost-plus pricing decisions, three key factors must be considered: the cost base, business volume level, and markup rate. This approach offers advantages such as simplified pricing, price stability, and leadership in pricing decisions. To streamline our model, the dissertation simplify the unit cost to solely consider the wholesale price.

The objective of our research is to investigate the relationship between the total sales volume of Category 1 vegetables in a day and cost-plus pricing. Dissertation treat the total sales volume of category 1 vegetables in a day as the dependent variable, with the cost-plus pricing of this category of vegetables serving as the independent variable. Initially, this dissertation employ Pearson correlation analysis to assess their relationship. The results indicate a statistically significant negative correlation at the 1% significance level, aligning with our daily observations that an increase in cost-plus pricing corresponds to a downward trend in the total sales volume of category 1 vegetables.

Subsequently, this dissertation aim to establish a multiple linear regression model between the total sales volume of category 1 vegetables and the cost-plus pricing of category 1 vegetables.

$$Q = \varepsilon + CS \tag{2}$$

Where ε and C represent the regression coefficients of the equation.

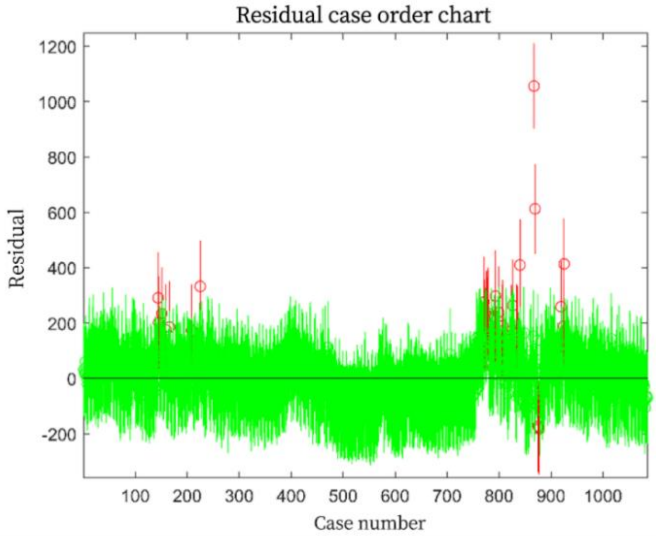

Figure 1. Residual plot

Table .1. Confidence intervals

Correlation coefficient R2	F value	Probability P corresponding to F	Estimation error variance
0.0295	32.8907	1.2637 e-08	7217.9585

The resulting linear correlation results are:

$$Q = 241.8798 - 8.0099\, S \tag{3}$$

As shown in Figure 1 and Table 1, observing the confidence interval, we find that the correlation coefficient R^2 is not 0, but it is indeed far away from 1, indicating poor linear correlation. Eliminating outliers from the residual plot introduces new outliers, and the integrity of the data worsens over time. The linear fitting results for other vegetable categories are similar to this case, with consistently poor linear correlation.

To gain deeper insights into the relationship between total vegetable sales and cost-plus pricing, The dissertation implemented a Support Vector Regression (SVR) model to enhance the accuracy of predicting vegetable demand [8]. The results are depicted in the Figure 2.

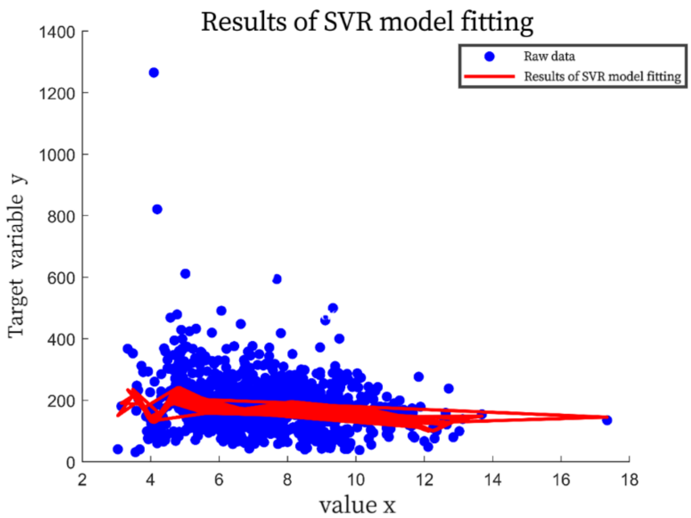

Figure 2. SVR fitting results

As shown in Figure 2, it is found that the scatter plots of total vegetable sales and cost-plus pricing are mostly clustered together, and the ribbon fitting results given by the SVR model better reflect this reality. However, if some outliers can be excluded, the fitting effect of the model will be further improved.

3. Vegetable Sales Trend Forecast Based on Seasonal Adjustment and LSTM

3.1. Sales Trends Based on Seasonal Index Forecast

Considering the significant influence of the day of the week observed in the initial question on the sales of various vegetable categories, and in the absence of legal holiday interference, the article employed the seasonal index prediction method with a window size of 7 days to forecast the sales of different vegetable categories over the next 7 days.

Seasonal index forecasting is an algorithm designed for analyzing time series data exhibiting periodic or seasonal fluctuations. This method primarily accounts for the seasonal components within time series data and predicts the value of a specific period in the future through statistical analysis of historical data [9].

In practical applications, seasonal indices are typically defined using the following mathematical expressions:

$$S_t = \frac{X_t}{M} \tag{4}$$

S_t is a seasonal index of time t, X_t is a raw observation of time t, and M is a moving average or other form of trend component.

With a 7-day window period, the dissertation calculate the average sales volume for each specific day of the week (e.g., the first day of the week, the second day of the week, etc.) within that window period. These averages serve as seasonal indices for the corresponding dates in the future. The prediction steps are as follows:

1. Data collection: Collect time series data over a sufficiently long period.
2. Window period selection: Select an appropriate window period to capture seasonal fluctuations in data. In this scenario, this article chose a 7-day window.
3. Seasonal index calculation: Average the data points in each window period to obtain the seasonal index of the day of each week.
4. Forecast model construction: use the seasonal index to predict data in a certain period in the future. This is often achieved by combining seasonal indices with secular trends and other components such as random error.
5. Model verification, the final forecast results of sales of each category are shown in the Figure 3, and a-f represents different vegetable types respectively.

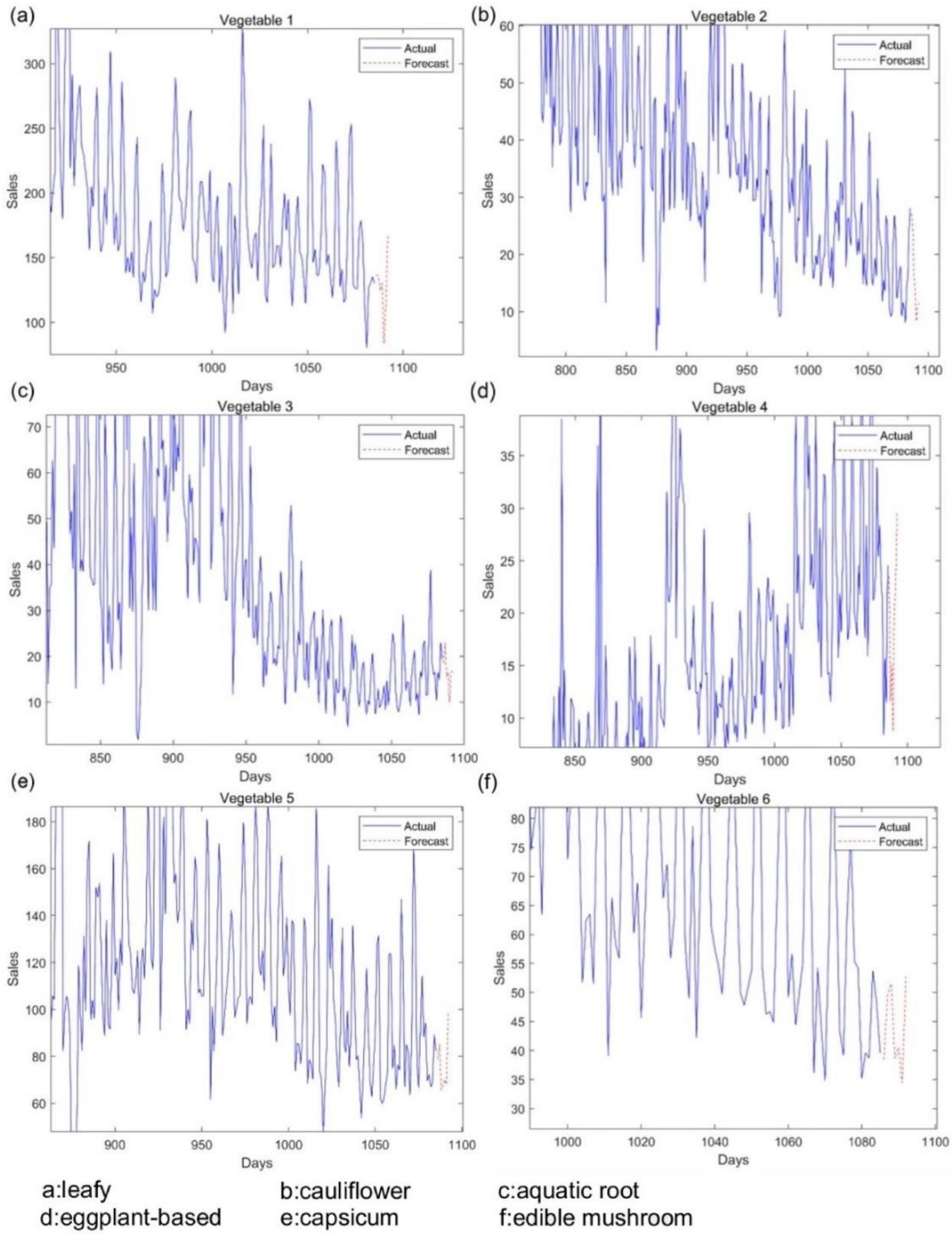

Figure 3. Line chart of forecast results of each vegetable category

As shown in Figure 1-3, it becomes evident that the seasonal index prediction algorithm aligns closely with previous patterns. Given that most varieties cannot be resold the following day if unsold, this dissertation consider the daily sales volume as the total daily replenishment volume to ensure it roughly matches the demand.

To maximize supermarket revenue, it is imperative not only to make accurate predictions regarding the sales of various vegetable categories in the next 7 days but also to formulate appropriate pricing strategies. The dissertation utilize the cost-profit margin as characteristic data for cost pricing. Conducting a Pearson correlation analysis on the daily sales volume and daily average cost-profit margin of vegetable category No. 1, the results obtained from SPSSPRO are depicted in the Table 2.

Table .2. Pearson correlation analysis between daily sales volume and daily average cost profit rate of vegetable category No. 1

	Sales	Cost margin
Sales	1(0.000***)	0.13(0.000***)
Cost margin	0.13(0.000***)	1(0.000***)

Note: * * *, * * and * represent the significance level of 1%, 5% and 10% respectively

The article observe a significant positive correlation between sales volume and cost margins, reaching the top 1% significance level, which may seem counterintuitive. Upon analyzing the data, the dissertation infer that this anomaly arises due to supermarkets' precise ability to forecast future sales and their adoption of effective pricing strategies. When forecasted sales volume is high, supermarkets tend to increase the expected cost-profit margin, whereas they reduce it when forecasted sales volume is low. This adjustment aims to mitigate the daily loss rate across various vegetable categories.

To devise a more optimized pricing strategy, the article extend seasonal index forecasts to encompass the cost-profit margin of various vegetable categories. The results, illustrated in the Figure 4, utilize label a-f to represent different vegetable types, respectively.

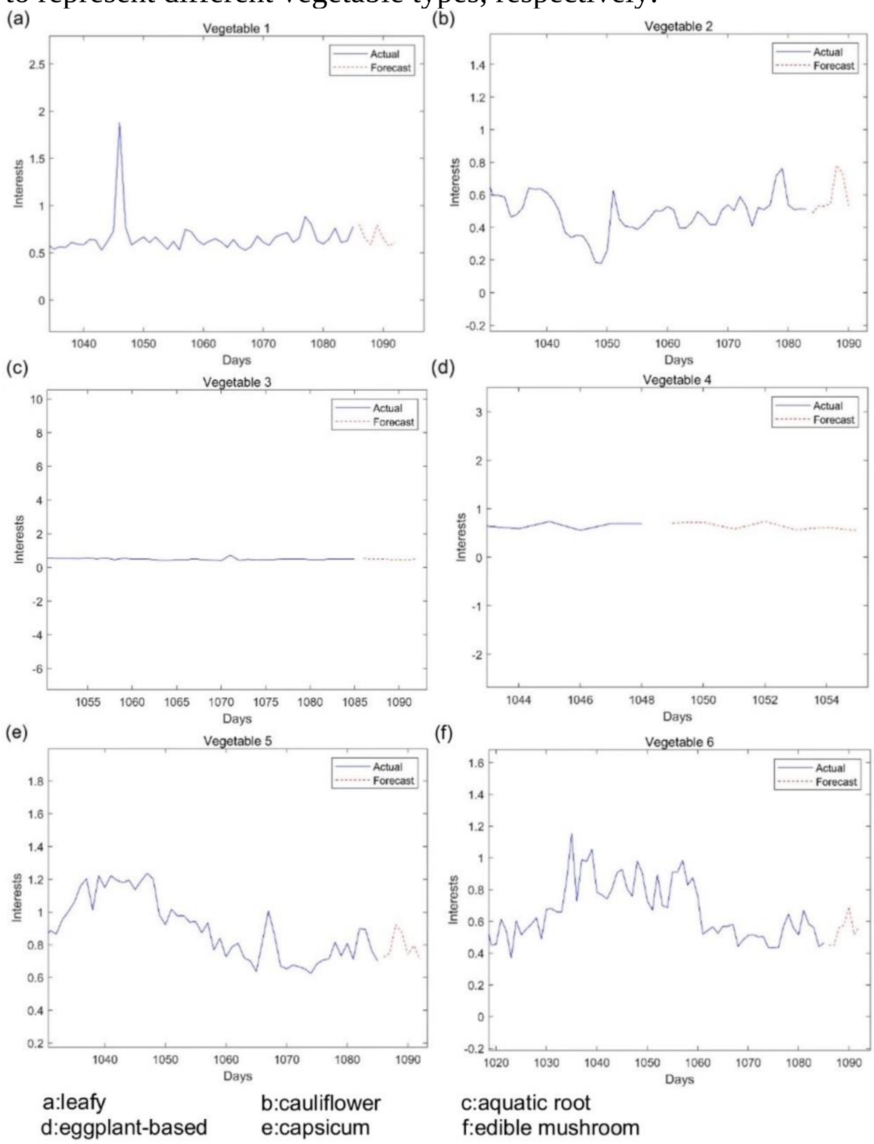

Figure 4. Seasonal index forecast chart of cost profit margin of vegetables of various

It can be observed that the forecast results are relatively stable and maintain good continuity with the trend of the previous period. We believe these forecasts are reliable, so we consider them one of the key factors in our July 1-7 pricing strategy.

Seasonal index forecasting is very beneficial for processing seasonal or cyclical time series data and can reduce the impact of seasonal or cyclical factors on the data. However, if the seasonal variation is not regular enough, the seasonal index method may not be suitable. Therefore, this dissertation plan to introduce long-term and short-term memory networks (LSTM) for comparative prediction to obtain a more accurate prediction model. The capability of Long Short-Term Memory (LSTM) networks to capture long-term dependencies has been acknowledged by numerous scholars and has found extensive applications in stock market prediction, wind power forecasting, short-term load forecasting, and sales forecasting. This indicates that LSTM is a highly potent tool for the prediction of agricultural product sales [10].

3.2. Time Series Prediction Based on LSTM (Long Short Time Memory Network) Model

To find the best market purchase and pricing decision-making plan, It is needed to predict the changing trend of sales volume in the next 7 days according to the existing time series data, to obtain the reference value of purchase volume. The sales volume of each category in supermarkets presents seasonal, dynamic, and cyclical characteristics. This dissertation can split a large amount of raw data into training sets and test sets, and then build a simple and flexible LSTM model for predicting future sales [11].

For the forecast of sales volume in the next 7 days, this dissertation will use 80% of the training set and 20% of the test set. To evaluate the accuracy of the prediction results, we will use the mean square error as an evaluation indicator:

$$EMS = \frac{1}{N}\sum_{i=1}^{n}(y_i - p_i)^2 \tag{5}$$

Taking the mosaic class as an example, set 7 as the moving step size, and make the LSTM based on the seasonal forecast and Adam's forecast. The neural network model is shown in Figure 5:

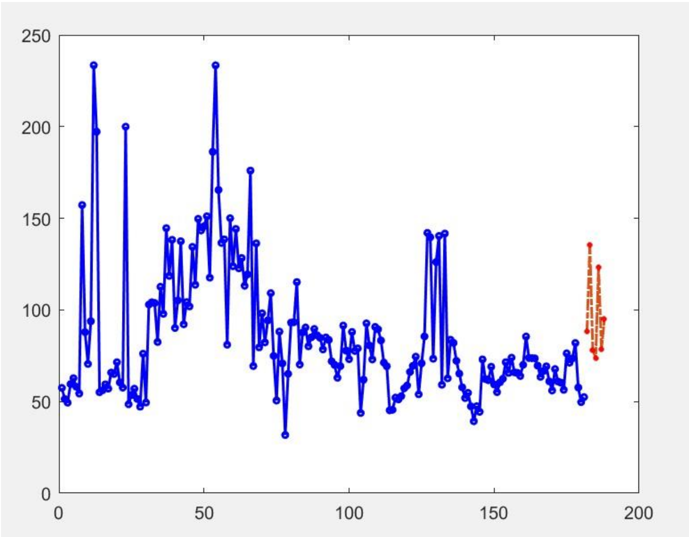

Figure 5. LSTM forecast based on the seasonal index and Adam optimization

4. Conclusions

To optimize the pricing strategy of fruits and vegetables, first of all, this dissertation use Pearson's correlation analysis method to explore the relationship between the total sales volume of category vegetables in one day and the cost-plus pricing of category vegetables. The results showed a negative correlation with significance levels reaching the first 1%. Although the article tried to establish linear regression equations, but their linearity was low. Subsequently, this article introduce a support vector regression (SVR) model to better reveal the relationship between category total sales and category vegetable cost-plus pricing. Subsequently, the dissertation constructed an improved long-term and short-term memory network (LSTM) model based on seasonal index prediction, predicted the sales volume in the next 7 days, and used it as the total daily replenishment volume for these 7 days. Finally, the article formulate a pricing strategy for the next 7 days based on these forecast results.

This paper provides an example of a research idea and framework applied to the fields related to fruit and vegetable market forecasting. These results provide decision makers with more reliable market forecasting information and help formulate more accurate pricing and replenishment strategies, thereby improving the operating efficiency and profit level of supermarkets that have commercial prospects and value.

References

[1] Jiang Dan. Research on consumer behavior analysis and market dynamic pricing model of agricultural products[D]. Northeast Agricultural University, 2014.

[2] Wang Xiaotong. Research on inventory control strategy of fresh fruit and vegetable products in supermarket chains based on DIT[D]. Southeast University,2020.

[3] Wang Ruogang. Problems and countermeasures of food safety management in agricultural supply chain[J]. China Food Industry, 2024,(01):95-97.

[4] Mao Lisha. Research on pricing strategy and production and marketing model of vegetable wholesale market under the perspective of supply chain[D]. Central South Forestry University of Science and Technology,2022.

[5] Lv Wang. Research on trend prediction of fresh vegetables sales based on improved SVM[D]. Anhui Agricultural University, 2021.

[6] Yang Ruiming,WANG Fuji,GUO Qianjin,et al. Research on fresh vegetable consumption of plantation structural adjustment in the context of agricultural supply side structural reform--Based on gray correlation analysis method[J]. Economist,2021,(10):37-39+42.

[7] Sun Kun, Research on demand forecasting of food supply chain based on machine learning [D]. North China Electric Power University (Beijing),2020.

[8] Zhang Qianqian. Research on vegetable demand forecasting system based on SVM [D]. Beijing Jiaotong University,2015.

[9] Nie Yuxuan. Research on automatic pricing and replenishment strategy of fresh commodities based on ARIMA prediction optimization model--taking vegetable commodities as an example[J]. Commercial Exhibition Economy,2024,(05):19-22.

[10] Cui Yunhao,Zhu Jun,Meng Hao. Application of neural network in sales forecasting of agricultural products[J]. Modern Agricultural Science and Technology,2021,(17):262-266.

[11] Cui Yunhao. Research on fresh vegetables sales prediction based on CNN-PSO-LSTM combination model[D]. Anhui Agricultural University,2022.