Prediction of Electric Load Neural Network Prediction Model for Big Data

Zifan Huang[1, #], Hongkui Zhong[2, #, *], Wenyu Chen[3, #], Jianhong Wang[3, #]

[1]Chang Chien College of Nantong University, Nantong University, Nantong, 226019, China

[2]School of Information Science and Technology, Nantong University, Nantong, 226019, China

[3]School of Science, Nantong University, Nantong, 226019, China

* Corresponding Author Email: td_ljf@163.com

[#]These authors contributed equally.

Abstract. Firstly, this article analyzes the distribution patterns and interrelationships of sales volume among different categories of vegetables and individual products. Based on the results of the stability test of the daily sales volume of six types of vegetables, the data of each type of vegetables is divided into 7 sub-data sets according to the week to explore the distribution patterns of vegetable sales volume and time. Subsequently, this article conducts a distribution test on 42 sub-data sets of the six types of vegetables, and finds that the Weibull distribution is the optimal fitting distribution. Spearman correlation coefficients are used to judge the correlation between the six types of vegetables based on scatter plots. The results show that there is weak correlation among different categories of vegetable products. An Apriori algorithm is used to analyze the correlation of 246 individual products, and 24 frequent itemsets are found. Then, a heatmap analysis is conducted on the individual products in each set, which shows strong correlation within the set. Finally, this article divides the daily sales volume of each vegetable category into two time scales, and uses the ARIMA(p,d,q) model and the ARIMA-LSTM model to predict the sales volume for the following week, which facilitates the construction of a sales volume-price model based on future predicted sales volume to predict the price of vegetables for the following week, and further analysis can be conducted.

Keywords: Weibull distributed, Apriori algorithm, ARIMA model, Expected utility theory.

1. Introduction

Under the policy background of supporting large chain supermarkets and farmers' markets, fresh supermarkets have ushered in a period of development opportunities. As an important sales channel, the scale of fresh supermarkets has exceeded trillions, and it is predicted that the sales scale of fresh supermarkets in China will reach 4 trillion in the future [1-2]. With the increase in residents' demand for the quality and freshness of vegetables, supermarkets usually replenish goods based on the historical sales and demand of each product every day. In fresh supermarkets, the preservation period of vegetable products is generally short, and their appearance deteriorates with the increase of sales time. Most varieties such as unexpired goods cannot be sold on the next day if not sold on the same day [3]. General supermarkets use the "cost-plus pricing" method, and usually sell damaged goods and products with deteriorated appearance through discounts[4]. Therefore, reasonable inventory management has a greater impact on the actual operating profit of supermarkets. To formulate reasonable inventory management, we must make replenishment decisions and pricing decisions based on reliable market demand analysis[5]. Therefore, this article takes the vegetable commodity information of a supermarket as the research object to predict the future sales of vegetables to facilitate subsequent decisions on replenishing and pricing vegetables.

2. Analysis of the sales distribution patterns and inter-relationships of various categories and individual products of vegetables

Firstly, this article conducts a stability test to judge the results of the time stability test, and analyzes the time series plot and unit root test. If the test results are unstable, it is necessary to refine the sales volume of each category from Monday to Sunday. If the test results are stable, a frequency bar chart can be directly fitted with the curve. The distribution law between categories and individual products is analyzed by fitting curve, and the parameters are estimated by maximum likelihood method, resulting in 42 distribution functions. Then, this article analyzes the scatter plot to obtain the nonlinear relationship between different categories, and uses Spearman correlation to solve the correlation between different categories. Finally, the mining algorithm is used to analyze the correlation of 246 individual products, and the purchase rules of individual products are obtained.

2.1. Analysis of monthly change patterns for each category

To explore the distribution law of sales volume of different categories and individual products of vegetables, this article firstly calculates the annual sales of 6 major categories of vegetables and draws a histogram of sales distribution for each category every year. By analyzing the histogram of annual sales statistics of different types of vegetables, it can be seen that leafy vegetables have the highest total annual sales for 3 years, while eggplants have the lowest, and from the overall distribution, it can be seen that the total annual sales of different categories are in descending order: leafy vegetables, peppers, mushrooms, cauliflower, aquatic plants, and eggplants, with a general trend of decreasing from high to low. Secondly, starting from the comparison of sales volume of different vegetable varieties in each quarter, a histogram comparing the sales volume of different vegetable varieties in each quarter is drawn, as shown in Fig.1. From the histogram comparing the sales volume of different vegetable varieties in each quarter, it can be seen that there may be a certain connection between the total sales volume of vegetables and time. However, it is not possible to clearly see the possible distribution law from the histogram of total sales volume. Therefore, starting from the total sales volume by month, day and year, we draw the total sales sum of each month, day and year for 6 major categories [6]. Taking cauliflower as an example, Fig. 2 plots the month-on-month change in cauliflower's monthly sales volume.

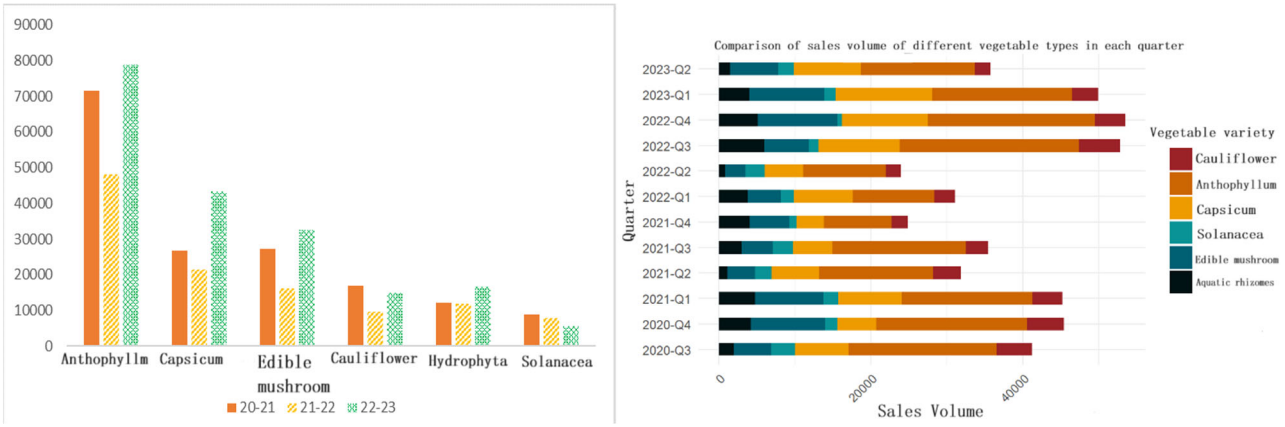

Fig. 1 Histogram comparing annual and quarterly sales volume statistics of different types of vegetables (left: annual, right: quarterly)

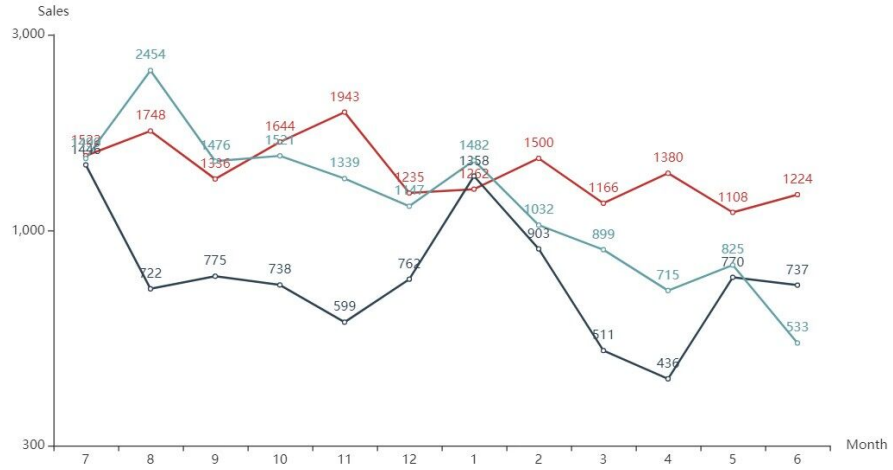

Fig. 2 Line chart showing the monthly sales trend of cauliflower category

From the monthly change line chart, it can be seen that the monthly sales of cauliflower in the three years vary significantly from year to year, and there is a large range of changes in each month during the three years, but there is no clear pattern. To explore in detail the possible distribution law of cauliflower sales and time, a time series plot of daily sales for each type of vegetable is drawn.

2.2. Stationarity test

The stationarity test of the daily sales volume of cauliflower shows that the P-values of the three types of τ statistics are significantly greater than the significance level ($\alpha = 0.05$), indicating that the data distribution of daily sales volume of cauliflower is not stationary. Then, this article draws the seasonal decomposition graph of cauliflower, as shown in Fig. 3, and finds that the daily sales volume of cauliflower shows a periodic distribution. Additionally, this article finds significant differences in sales volume for different weeks, and accordingly refines the sales volume of each type of vegetable into weekly sales volume, i.e., splitting them into 7 sub-sequences according to weekly patterns.

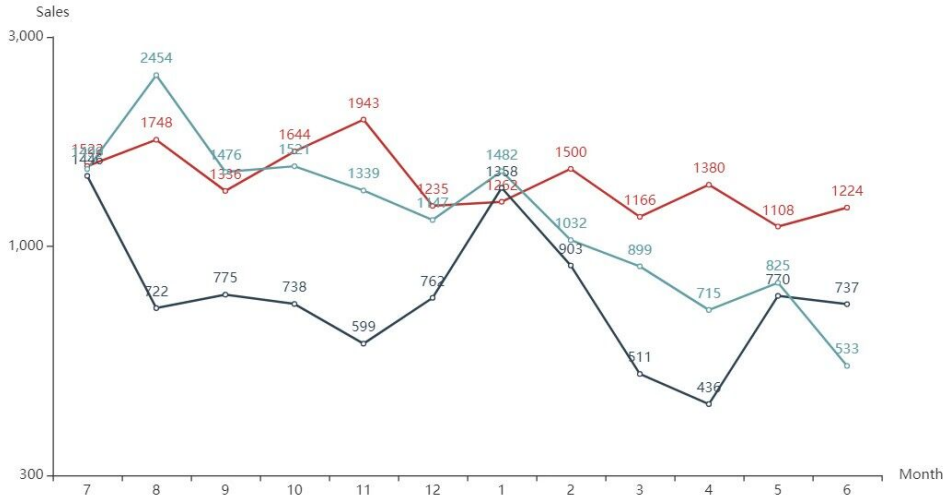

Fig. 3 Seasonal decomposition chart of cauliflower category

2.3. Explore the distribution patterns of each category

Taking cauliflower as an example, the frequency distribution histogram of the total sales from Monday to Friday is plotted. The frequency distribution histogram of cauliflower sales on Monday and Sunday is shown in Fig. 4. Based on the histogram of total cauliflower sales by week, it can be seen that the overall distribution follows a bell shape, with higher sales in the middle and lower sales on both sides. Therefore, it is likely that the distribution of cauliflower sales follows a specific distribution, such as normal distribution, Weibull distribution, or t-distribution.

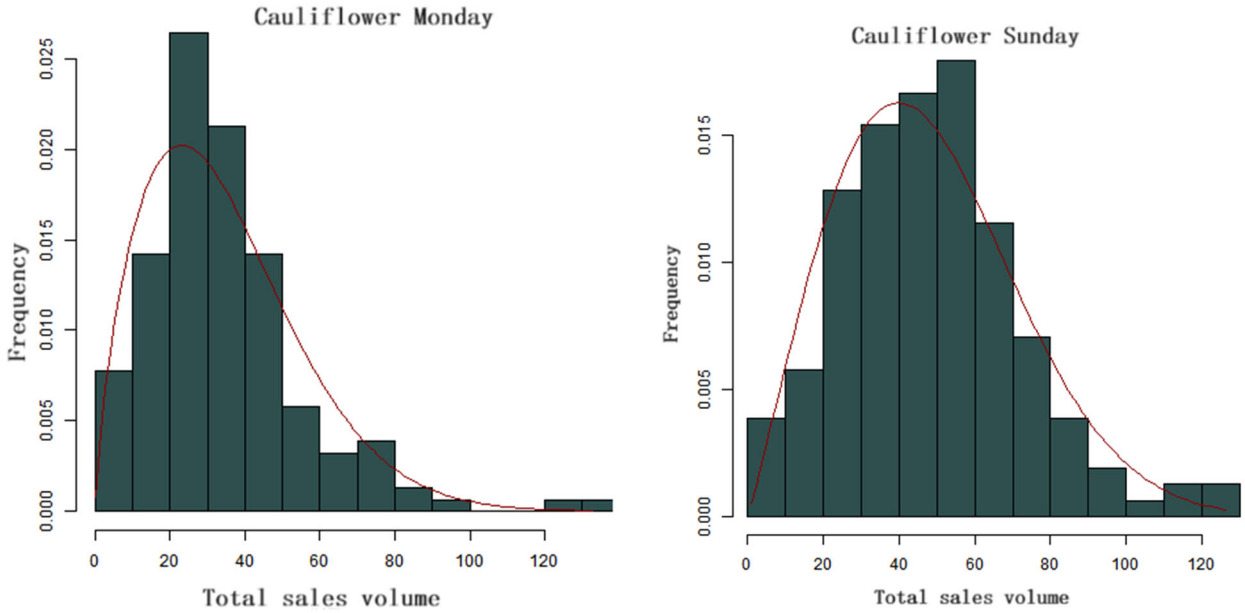

Fig. 4 Histogram comparing the sales volume of cauliflower on Monday and Sunday

To further explore the distribution law, taking the total sales of cauliflower on Monday as an example, the normal distribution, Weibull distribution, t-distribution and lognormal distribution were tested for the relevant data and the relevant P-P plots were plotted as shown in Fig. 5 and Fig.6. From the four P-P plots, it can be seen that the data can follow the P-P line of the three distribution theories, but the best fit for the data is the theoretical P-P line of the Weibull distribution. Therefore, it can be considered that the total sales of cauliflower on Monday follow the Weibull distribution. The maximum likelihood estimation method was used to estimate the relevant parameters of the Weibull distribution and fit the relevant curve. The three distributions were tested for each of the 6 different varieties of vegetables and their corresponding 7 sub-data from Monday to Sunday, and the relevant P-P plots were plotted. The conclusion is that the 42 sets of data corresponding to the 6 categories all follow the Weibull distribution.

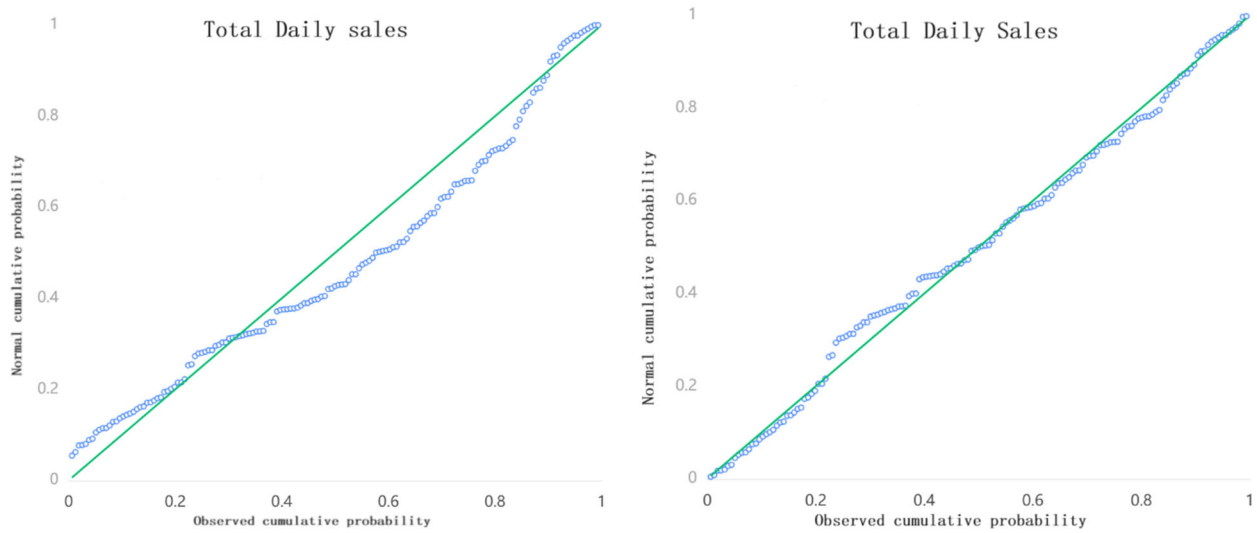

Fig. 5 P-P plot for normal distribution and Weibull distribution validation

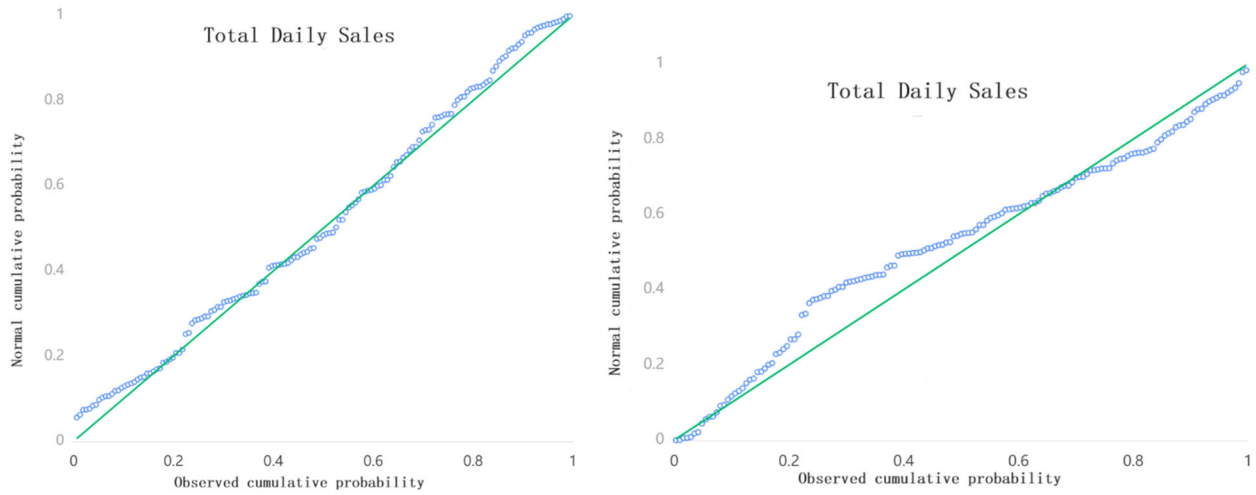

Fig. 6 P-P plot for t-distribution and lognormal distribution validation

2.4. Correlation analysis between different categories

To further analyze the possible correlation relationships between different categories or individual products of vegetable commodities, this article first analyzes whether there is a correlation between different categories of vegetable commodities. By drawing scatter plots of the sales of six categories of vegetables, it can be seen that the scatter plots between each pair of categories cannot characterize the linear correlation between different vegetable categories. In other words, there are complex non-linear relationships between each type of vegetable. Therefore, the Spearman correlation analysis method is used to analyze the non-linear relationships between different categories of vegetables. Fig.7. shows scatter plots between three pairs of vegetables.

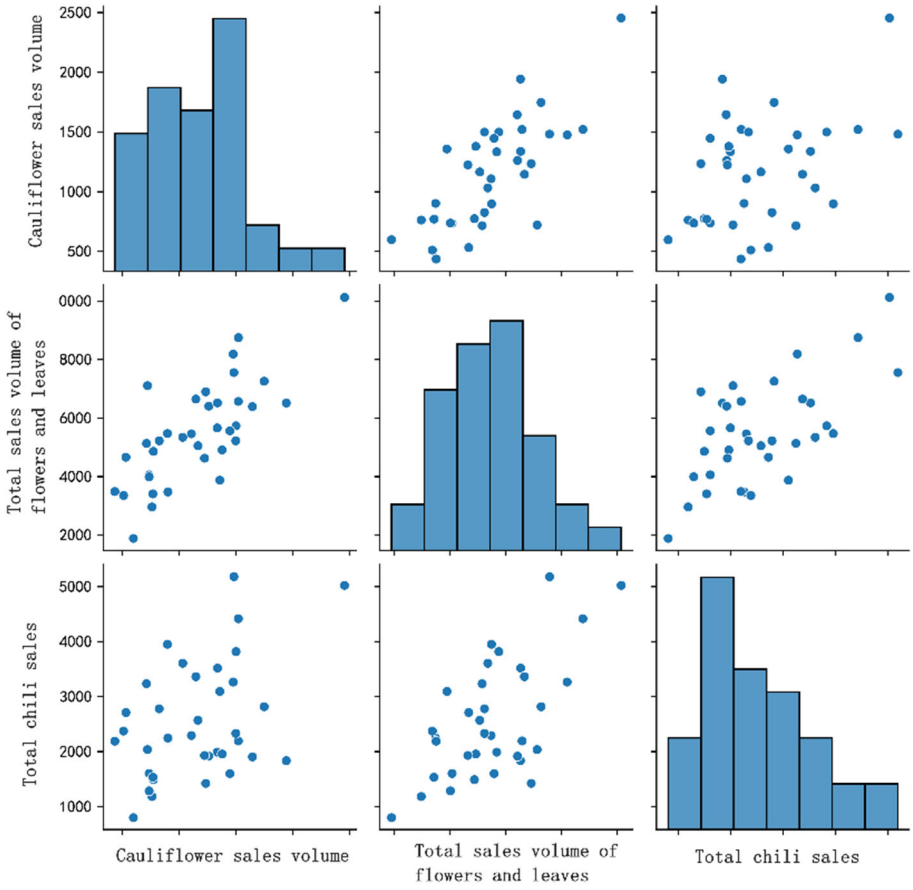

Fig. 7 Scatter plot of three categories pairwise

For a sample of size n, the n original data are converted into rank data, and the correlation coefficient ρ is calculated as shown in formula (1):

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \tag{1}$$

x_i and y_i represent the ranks of x and y, respectively, while $\bar{x}$ and $\bar{y}$ represent the average ranks.

Using SPSSPRO, Spearman correlation analysis was conducted on the sales relationship of the six major categories of vegetables after pre-processing, and the Spearman correlation heatmap is shown in Fig.8.

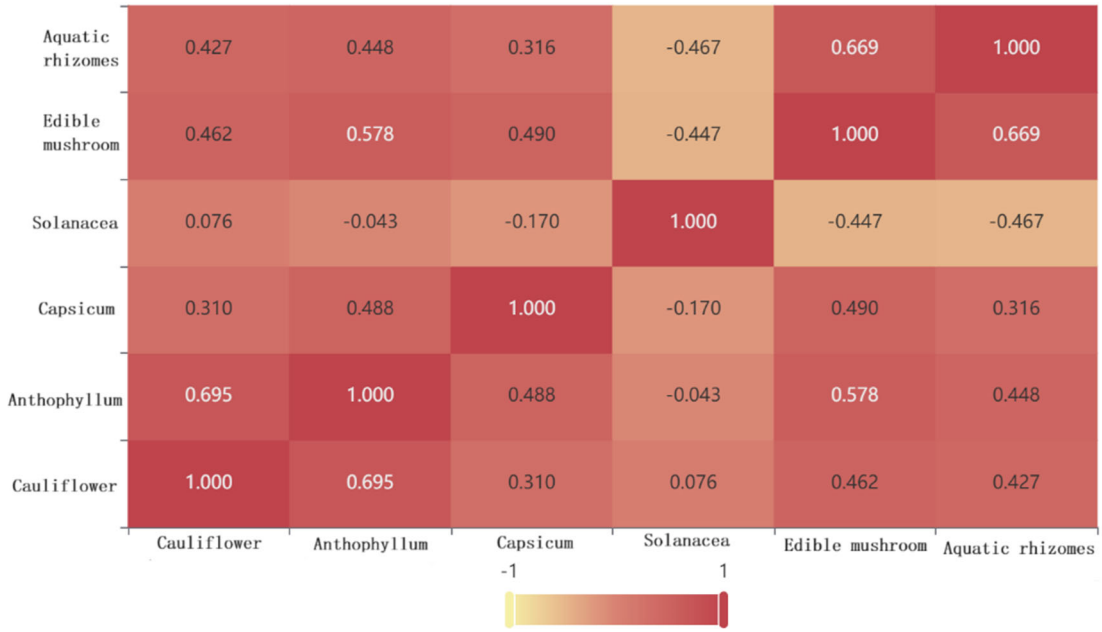

Fig. 8 Heat map of correlation coefficients between six categories

By analyzing the heatmap of Spearman correlation coefficients, it can be seen that the maximum Spearman correlation coefficient between all categories of vegetable commodities is 0.695 between leafy vegetables and cauliflower, and except for a correlation coefficient value of 0.669 between aquatic root vegetables and edible fungi, most of the other values are below 0.5, indicating that most of the different categories of vegetable commodities do not have strong non-linear correlation with each other, i.e. there is no strong correlation between them. Therefore, it can be roughly concluded that there is no correlation between different categories of vegetable commodities, i.e. they are mutually independent.

2.5. Correlation analysis between different individual products

Considering that there are many individual data items among different categories of vegetable commodities, with a wide variety of data types and complex data quantities, for the analysis of the correlation between different individual products, this article proposes the application of Apriori algorithm for data mining of associated product combinations. Apriori algorithm is a commonly used method for data association rule mining. The core idea of this algorithm is to calculate the support degree of itemsets by scanning the database multiple times to discover all frequent itemsets and generate association rules.

Firstly, this article performs correlation analysis on the individual products in the frequent itemset. Using SPSSPRO, Spearman correlation analysis was conducted on the selected individual products, and the Spearman correlation heatmap is shown in Fig.9.



Fig. 9 Association result plot of six different individual products

From Fig.9, it can be seen that the correlation values between Yunnan oilseed rap and green string peppers, green eggplants, and yellow cabbages are all above 0.6, indicating strong correlation between each pair of individual products. Therefore, it can be concluded that the pairs of individual products mentioned above are potential sales combinations.

3. Establishment and analysis of time series prediction model for vegetable sales

Considering that supermarkets need to plan replenishment based on categories, this article first uses the ARIMA model to predict the daily replenishment volume of each vegetable category for the next week based on daily sales. Through data analysis, each category is divided into seven weeks, and the ARIMA model and ARIMA-LSTM model are used to predict the daily sales from Monday to Sunday for each category. The prediction results are then analyzed and compared to facilitate the analysis of the daily replenishment volume of each vegetable category for the next week and the corresponding pricing strategy.

3.1. Time series prediction of future weekly price

Due to the significant trend, periodicity, and diversity observed in the sales volume sequences of each vegetable category, which indicates non-stationary time series, the ARIMA(p, d, q) model is considered for sales volume prediction in this article[7-8]. For a few stationary time series, we consider several commonly used models including AR(p) model, MA(q) model, and ARMA(p, q) model, as these time series lack significant trend features. The basic principle of model identification is shown in Table.1.

Table 1. Model identification principle

Autocorrelation function	Partial autocorrelation function	Model category
Trailing p-order truncation	p-order truncation	AR(p)
q-order truncation	Trailing p-order truncation	MA(q)
Trailing p-order truncation	Trailing p-order truncation	ARMA(p, q)

3.2. Prediction of future weekly price based on three consecutive years of data

Firstly, this article summarizes the daily sales of each individual product within each vegetable category, and treats each category's daily sales sample as an initial time series after data processing. The time series for six different categories are plotted separately. As the time series of the aquatic root vegetables and eggplants display a periodic distribution pattern, these two sequences are initially identified as seasonal non-stationary time series, and further verified by combining the time series images and unit root test-ADF method. ARIMA models are then established for the six time series respectively, and data fitting is performed. The residual sequence is then tested using the LB test, and the test statistic value is greater than 0.05 for all six models, indicating that the residual sequence is a white noise sequence. This indicates that these six models are significantly effective. Fig.10 displays the model prediction results for cauliflower and leafy vegetables.

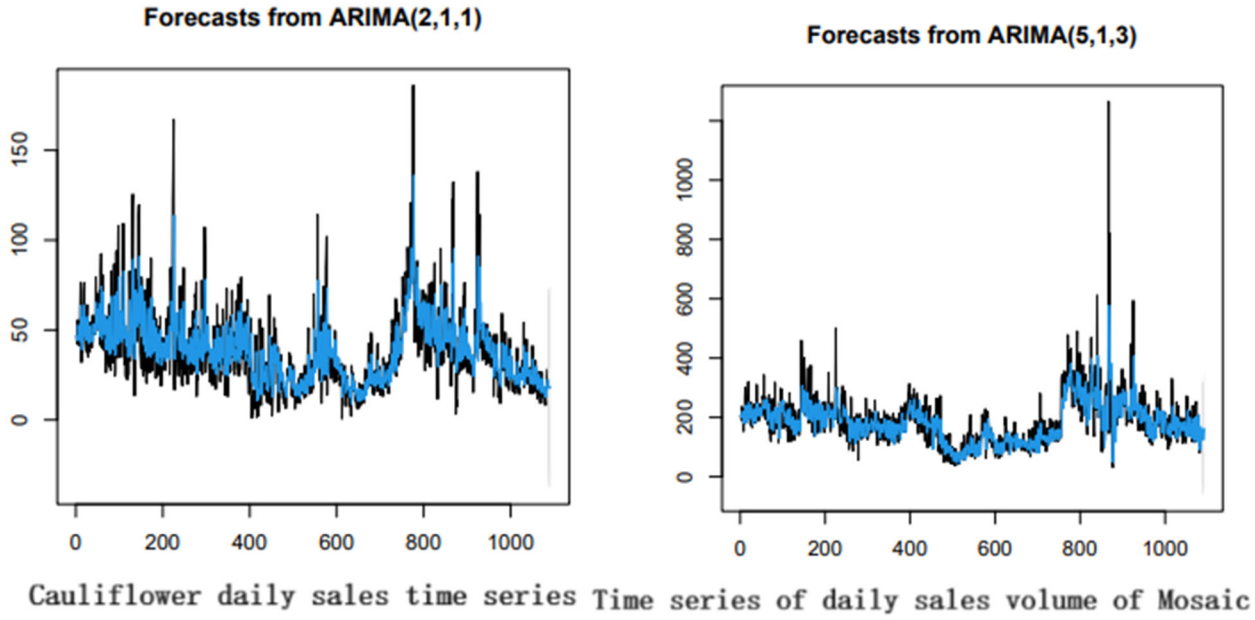

Fig. 10 Time series plot of daily sales prediction models for cauliflower and leafy vegetables categories

3.3. Prediction of future weekly price based on data divided by week

As shown in Figure 10, the daily sales data of each vegetable category exhibits an obvious cyclical pattern, and significant differences in the sales quantity of different weeks exist. Therefore, the sales volume of each vegetable category is refined into 42 time series from Monday to Sunday, denoted as . Firstly, the time series of the 42 time series are plotted, and the stationarity of the sequence is verified by combining the time series image and unit root test. Through analysis, it is found that the daily sales volume time series shows obvious cyclical patterns for vegetable categories such as aquatic root vegetables. After re-division, there is no significant cyclical feature. Fig.11 plots the sales volume time series of Tuesday and Sunday for aquatic root vegetables, eliminating the periodicity. For stationary time series, this article adopts the ARIMA(p,q) model for modeling; for non-stationary time series, the sequence is first smoothened through d-order difference, and then white noise testing is used to confirm whether to fit the ARIMA(p,q) model. Subsequently, white noise testing is performed again on the residual sequence generated by fitting the model. Subsequently, this article establishes 42 ARIMA models and verifies the fitting effect through AIC and MASE indicators. The results show that the model fitting effect is good and the fitting prediction image is generated.

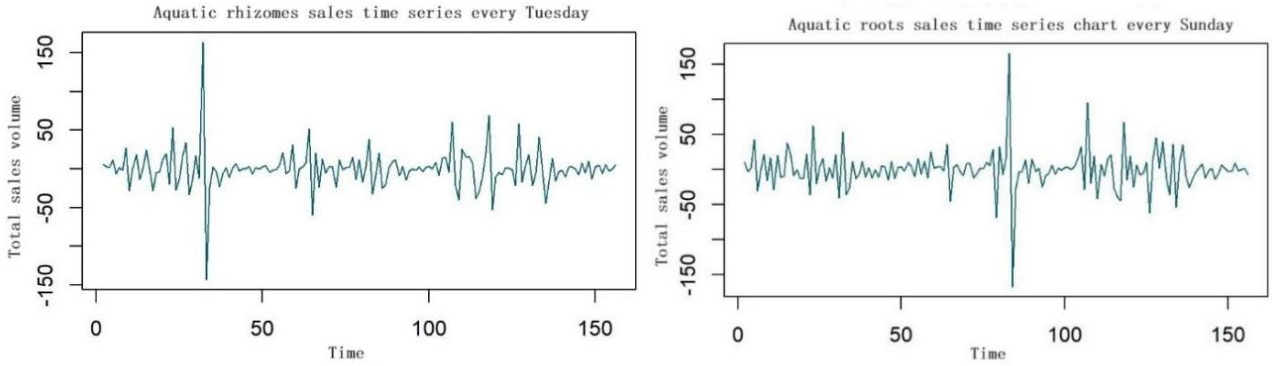

Fig. 11 Time series plot of sales on Tuesday and Sunday for aquatic root and stem vegetables categories

To ensure the accuracy of the prediction results, we compared and analyzed the prediction results of the ARIMA models with two different time series partitioning methods. Through the comparison results, this article determines that the built ARIMA model can better predict sales volume, and both the model accuracy and precision are high. Therefore, this article finally selects the 42 ARIMA model prediction results divided by week to determine the selling price, as their prediction trends are relatively consistent.

3.4. The solution of ARIMA-LSTM model

The article builds ARIMA-LSTM models for seven time sub-sequences of each of the six major categories of vegetables [9-10]. Taking the time sub-sequence data of cauliflower on Monday as an example, the article constructs an ARIMA-LSTM model. Firstly, the article performs data preprocessing on the time series, including difference calculation to stabilize the data and perform ARIMA modeling to obtain the predicted results. Secondly, the article divides the time series data into a training set and a test set, establishes an LSTM neural network model, uses the training set to train the LSTM model, and obtains the loss graph of the training process model as shown in Figure 12. Finally, the article uses the trained LSTM model to predict the test set and obtains the predicted results of the LSTM model as shown in Fig 13.

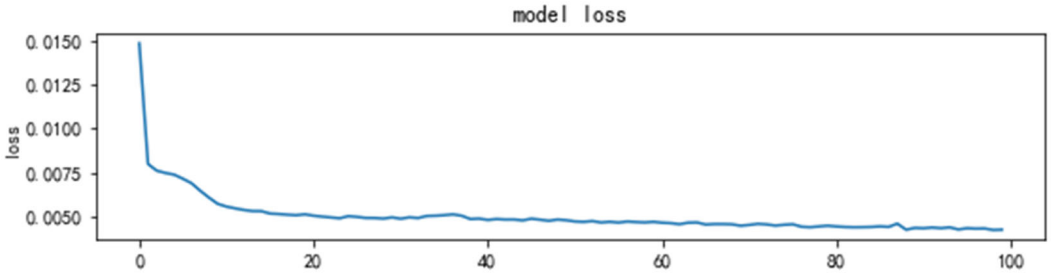

Fig. 12 Loss plot for training data

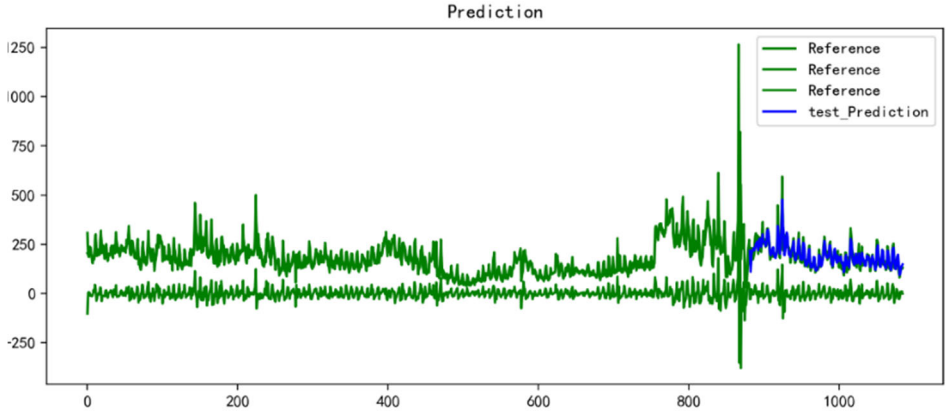

Fig. 13 Prediction results of the LSTM model on the test set

4. Conclusions

This article fully considers the distribution patterns and inter-relationships of various categories and individual products of vegetable sales, and ultimately determines the Weibull distribution as the optimal fitting distribution. At the same time, this article gradually explores the possible sales relationships between different categories and individual products of vegetables in a step-by-step approach from major categories to minor categories. For the six major categories of vegetables, the Spearman correlation coefficient is used to measure the correlation between the six categories of vegetables. For each individual vegetable product, a correlation mining model is constructed, and the Apriori algorithm is used to analyze the correlation between all vegetable products. This quickly and visually identifies potential demand relationships between different products, finds strongly correlated pairs of products, and uses them as the most likely sales combinations to lay a foundation for future product analysis and selection in sales combinations. Next, this article divides the daily sales of each vegetable category into two time scales, fully considering the seasonal and distribution trends of data time distribution. The ARIMA(p,d,q) model and ARIMA-LSTM model are used to predict future weekly sales for easy construction of a sales-pricing model based on future pre-sales for predicting future weekly vegetable pricing, and further analysis can be conducted.

References

[1] Wang Zengfei, Wang Xiaodong, Zhao Anping et al. Market situation and future prospects of Beijing vegetable market in the first half of 2023 [J]. Agricultural Outlook, 2023, 19 (07): 18-23.

[2] Zhang Jing, Liu Jifang, Wu Jianzhai et al. Analysis on operation of vegetable market in 2021 and prospect in 2022 [J]. Chinese Vegetables, 2022 (01): 1-8.

[3] Li Xiaolulu, Zhou Shuguang. Study on development issues of fresh supermarket retail industry in China [J]. Journal of Business Economics, 2021, (23): 35-37.

[4] Ma Xiaoli. Research on inventory management optimization of fresh products in Y chain supermarket [D]. Hebei University of Science and Technology, 2022.

[5] Shen Chen, Zhao Feng, Sun Jiabo et al. Reflections and suggestions on vegetable market regulation in China [J]. Chinese Vegetables, 2022 (10): 14-19.

[6] Gao Yidan, Zong Yixiang. Analysis on vegetable price trend in the first quarter of 2020 in Hebei Province and prediction for the future [J]. Vegetables, 2020 (06): 71-77.

[7] Mao Lisha. Research on pricing strategy and production and marketing model of vegetable wholesale market from the perspective of supply chain [D]. Central South University of Forestry and Technology, 2023.

[8] Tian Dong, Wei Xinhua, Wang Yue et al. Temperature prediction in edible fungi greenhouse based on MA-ARIMA-GASVR[J]. Transactions of the Chinese Society of Agricultural Engineering, 2020, 36(03): 190-197.

[9] Cao Xinyue, He Chunlin, and Cui Mengtian. Urban vegetable price fluctuation patterns and prediction based on the X12-ARIMA and LSTM combined model. Journal of Southwest University for Nationalities (Natural Science Edition), 2021, 47(04): 418-425.

[10] Li Yang, Du Ruishan, Cheng Yongchang. Study on crude oil production time series prediction based on ARIMA-LSTM combined model [J]. Mathematics in Practice and Theory, 2022, 52 (06): 40-48.