Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

Research on Sales Decision of Vegetable Products Based on Time Series Analysis

Wenting Ma *, Jiawei Yang, Haitao Zhao

School of Computer and Information Engineering, Tianjin Normal University, Tianjin, China, 300387

* Corresponding author: 16622907629@163.com

Abstract. This article studies the automatic pricing and replenishment decision-making of vegetable products. Through factor analysis and time series analysis, analyze the inherent laws, and combine mathematical programming with genetic algorithm to make sales decisions for vegetable products. Firstly, the data is preprocessed using time series analysis to predict and fill in missing values from September 12, 2022 to October 14, 2022, when the data is empty. Secondly, the relationship between vegetable types was analyzed. JB and Q_Q tests were conducted on the daily sales of each category, and it was found that the data did not follow a normal distribution. Therefore, this article uses Spearman correlation coefficient to analyze the correlation between different categories and finds that the flavor, taste, or cooking characteristics of edible mushrooms and water rhizomes are complementary and can be used in combination. To analyze the relationship between the sales volume and price of vegetable varieties, this article uses linear regression analysis to find that there is a negative correlation between sales volume and price. Finally, considering the shortcomings of time series analysis and factor analysis models, combined with practical considerations, the data to be collected includes competitor data, seasonal sales data, market trends and consumer insights data, inventory data, weather data, promotion and discount data, purchasing power data, customer feedback and demand data, vegetable supply chain data, and loss rate data.

Keywords: Vegetable Sales; Factor Analysis; ARIMA; Genetic Algorithm.

1. Introduction

In actual production and life, the quality and appearance of vegetable products are easily affected by factors such as time, which in turn affects sales. Supermarkets need to replenish vegetables in a timely manner based on their shelf life and changes in appearance [1]. At the same time, they need to conduct market demand analysis and sell products with poor phase change through discounts and other methods [2].

Therefore, the replenishment and pricing decisions of vegetable products are very important for the operation of supermarkets. Supermarkets need to comprehensively consider factors such as shelf life, product appearance, market demand, and supply situation to avoid inventory backlog and losses [3]. They need to analyze the relationship between individual products, calculate the total sales volume of individual products, and create statistical charts for analysis. Analyzing the relationship between vegetable categories and conducting JB and Q_Q tests on the daily sales volume of each category, it was found that the data did not follow a normal distribution [4]. Therefore, this article uses Spearman correlation coefficient to analyze the correlation between different categories.

By using factor analysis to discover hidden factors between different categories, and through time series analysis, the distribution pattern of daily sales volume of categories is analyzed. At the same time, mean, maximum, minimum, standard deviation, skewness, and kurtosis are introduced to describe statistical measures and analyze the statistical pattern of daily sales volume of categories.

2. Spearman correlation coefficient analysis of the relationship between categories

The Spearman correlation coefficient is a non-parametric statistic used to measure the monotonic relationship between two variables [5]. It is a generalization of the Pearson correlation coefficient, which can be used to evaluate the hierarchical relationship between two variables, not limited to linear relationships. The JB test and Q_Q test was conducted on the daily sales volume of the category, and it was found that the data did not follow a normal distribution [6]. Therefore, this article uses Spearman correlation coefficient to analyze the correlation between different categories.

$$r_s = \frac{\sum (R_X - \bar{R}_X)(R_Y - \bar{R}_Y)}{\sqrt{\sum (R_X - \bar{R}_X)^2 \sum (R_Y - \bar{R}_Y)^2}} = \frac{\sum R_X R_Y - \frac{(\sum R_X)(\sum R_Y)}{n}}{\sqrt{\left(\sum R_X^2 - \frac{(\sum R_X)^2}{n}\right)\left(\sum R_Y^2 - \frac{(\sum R_Y)^2}{n}\right)}} \tag{1}$$

2.1. Spearman correlation coefficient

2.1.1. Normality test.

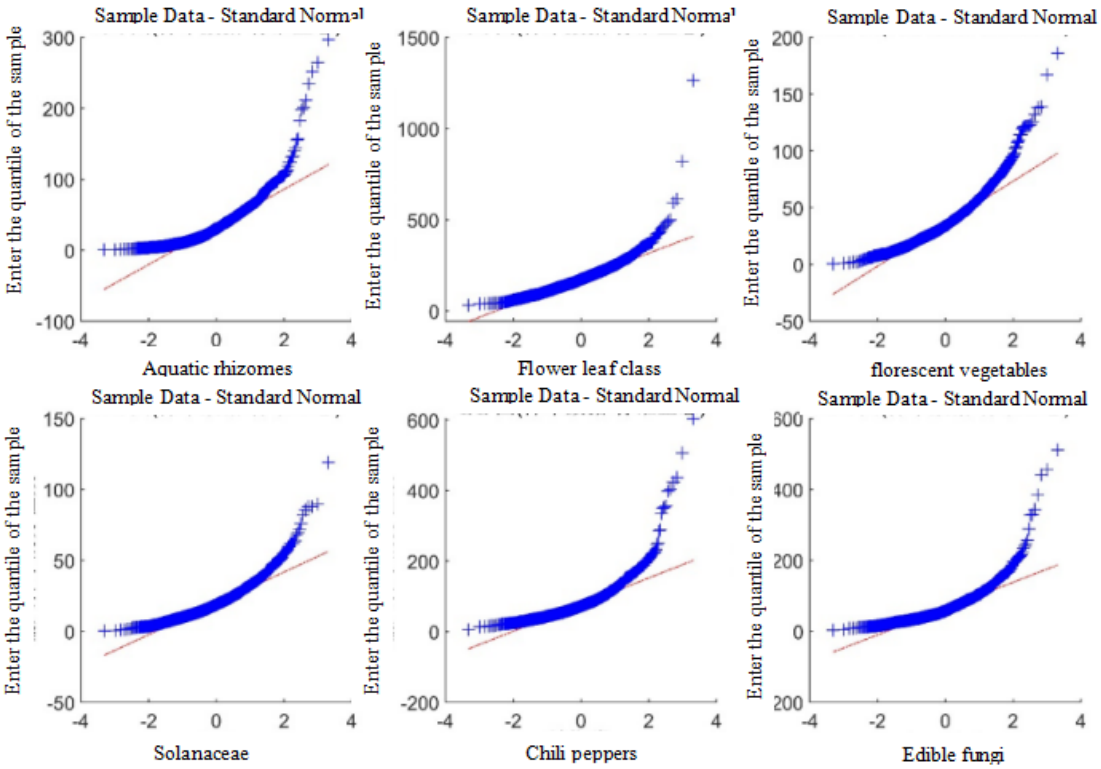

Figure 1. Normal distribution test

From Fig 1, the scatter points of the six variables deviate more from the diagonal distribution, indicating that the two variables do not conform to the normal distribution. Moreover, the JB test shows that at the 95% confidence level, these data do not conform to the normal distribution.

2.2. Monotonicity judgment

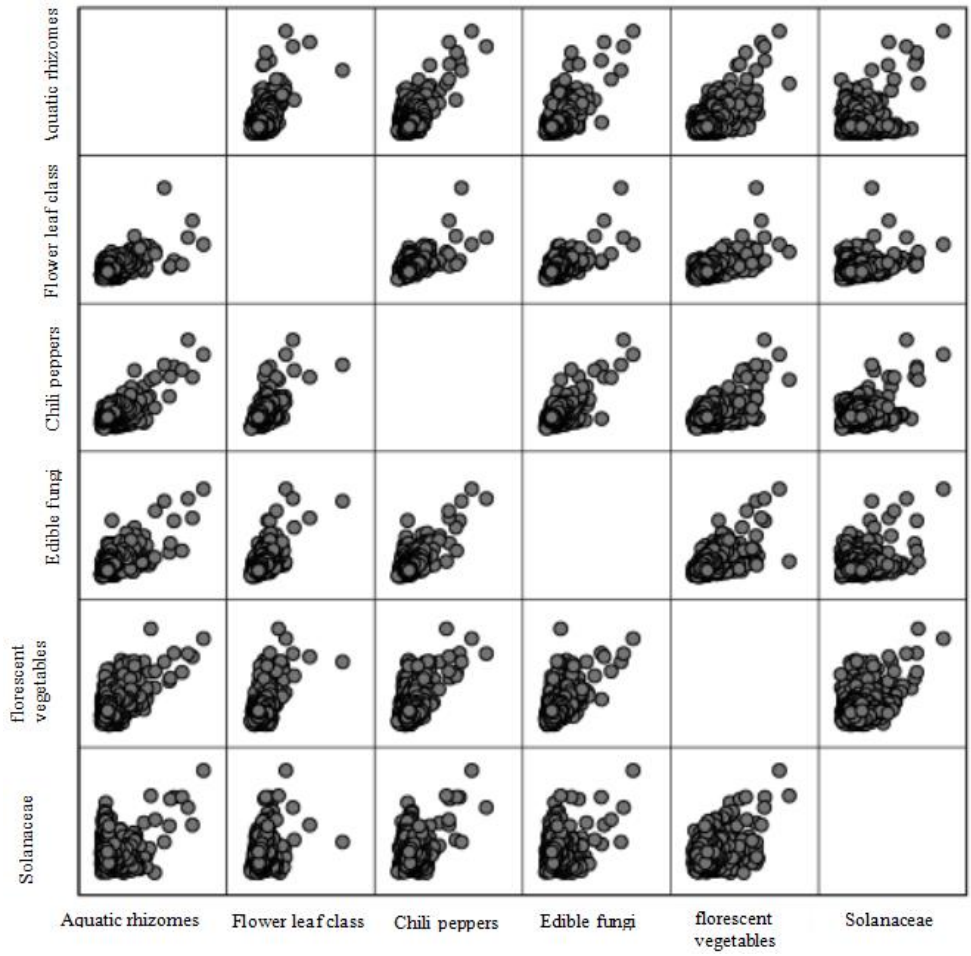

Figure 2. Scatter plot between variables

From Fig 2, there is a monotonic relationship between variables. In summary, the Spearman correlation coefficient can be used in this example.

2.3. Factor analysis identifies potential influencing factors of vegetable categories

Factor analysis is a commonly used multivariate statistical method used to explore and analyze the potential structure between multiple observed variables [7]. When analyzing the daily sales volume of a category, factor analysis is used to identify potential influencing factors or construct comprehensive indicators.

The general model of factor analysis:

$$\begin{cases} x_1 = u_1 + a_{11} f_1 + a_{12} f_2 + \cdots + a_{1m} f_m + \varepsilon_1 \\ x_2 = u_2 + a_{21} f_1 + a_{22} f_2 + \cdots + a_{2m} f_m + \varepsilon_2 \\ \quad \vdots \\ x_p = u_p + a_{p1} f_1 + a_{p2} f_2 + \cdots + a_{pm} f_m + \varepsilon_p \end{cases} \tag{2}$$

(1) Applicable condition determination

KMO test and Bartlett's sphericity test

Table 1. KMO and Bartley's test

name	numerical value
KMO sampling suitability quantity	0.851
Approximate chi square	3049.755
freedom	15
significance	0.000

KMO test standard: KMO>0.9, very suitable for factor analysis; 0.8<KMO<0.9, suitable; 0.6<KMO<0.7, not very suitable. According to Table 1, the KMO value is equal to 0.851, indicating that the data is suitable for factor analysis; The p-value of the Bartlett's sphericity test is 0.000, which is less than 0.05, indicating that we reject the null hypothesis at a 95% confidence level, that is, we believe the data is suitable for factor analysis.

(2) Determine the number of factors

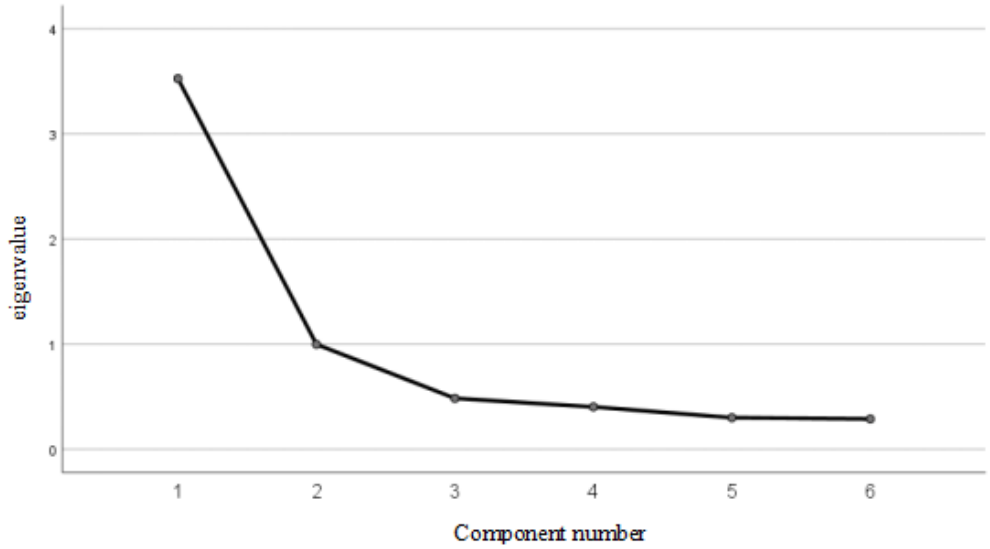

Figure 3. Crushed Stone Diagram

From Fig 3, the changes in the eigenvalues corresponding to the first two factors are relatively steep. Starting from the third factor, the changes in the eigenvalues are relatively flat. Therefore, we should choose two factors for analysis.

2.4. Recalculate the number of adjustment factors

Table 2. Component matrix a after rotation

Vegetable category name	Component 1	Component 2
Aquatic rhizomes	0.885	-0.069
Flower leaf class	0.797	0.280
Chili peppers	0.822	0.233
edible fungi	0.875	-0.005
florescent vegetables	0.706	0.382
Solanaceae	0.083	0.963

2.5. Result analysis and conclusion

After conducting factor analysis on vegetable sales data, Table 2 was obtained using the extraction and rotation methods of principal component analysis and Caesar normalization maximum variance method. After three iterations of convergence, we obtained the rotated component matrix a.

Based on the value of component matrix a, this article can provide a reasonable explanation for two factors. Based on the value of component matrix a, we can obtain the following explanation: Factor 1: aquatic rhizomes, flowers and leaves, chili peppers, edible mushrooms, and cauliflower with higher average sales have higher positive and negative load values [8]. This may mean that these categories have strong correlation on factor 1. Therefore, we can interpret factor 1 as the "green vegetable factor", which represents the degree of preference or specific taste requirements for these categories. Factor 2: Both eggplants and those with higher average sales have higher positive and negative load values, while other categories have lower load values. This indicates that eggplants have a significant impact on factor 2. Therefore, factor 2 can be interpreted as the "tomato factor", representing the degree of consumer preference for eggplant products.

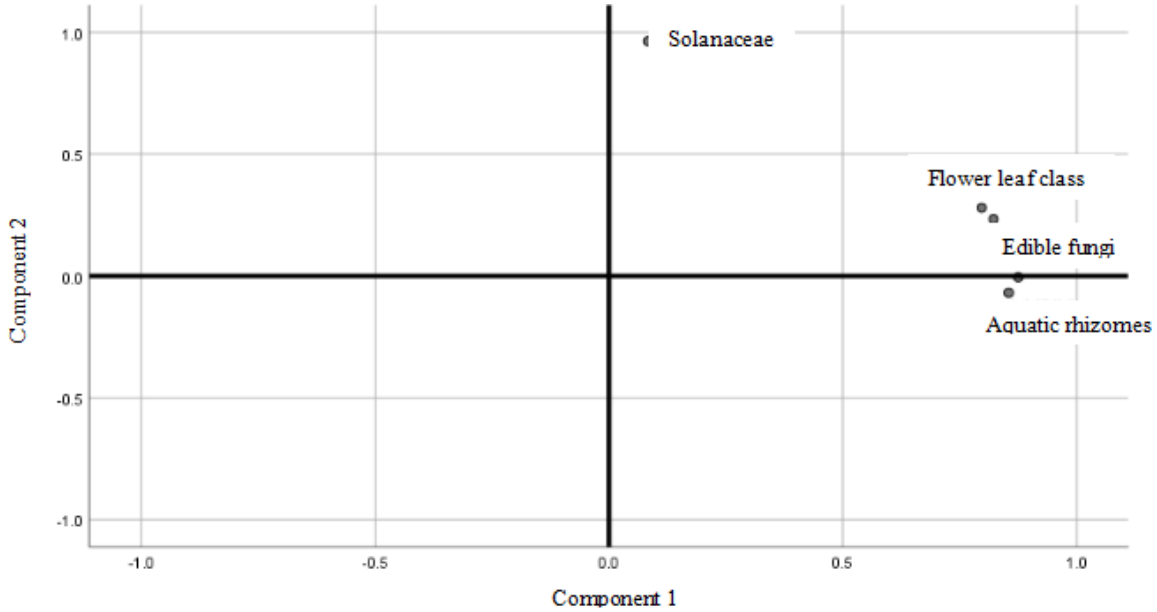

Figure 4. Scatter plot of factor load after rotation

Based on the two columns of data in the "rotated component matrix", the information observed in Fig 4 is consistent with the "rotated component matrix"the information obtained from the matrix is consistent.

2.6. The distribution pattern of sales volume in different categories

This article studies the distribution patterns of various categories of vegetables, and conducts statistical analysis on six types of vegetables, including mean, maximum, minimum, standard deviation, skewness, and kurtosis.

(1) Skewness

Skewness is used to measure the symmetry of probability distribution. A positive skewed distribution indicates that the tail of the distribution extends to the right, a negative skewed distribution indicates that the tail of the distribution extends to the left, and a skewness of 0 indicates that the distribution is relatively symmetrical [9].

The calculation formula is as follows:

$$\alpha = \mathrm{E}[(\frac{X-u}{\sigma})^3] \tag{3}$$

Among them μ It's the mean, δ Is the standard deviation.

(2) Kurtosis

Kurtosis is used to measure the kurtosis of a probability distribution, that is, the sharpness of the data. The kurtosis of a normal distribution is 3. If the kurtosis is greater than 3, the distribution exhibits a sharp peak shape, while if the kurtosis is less than 3, the distribution is relatively flat.

The calculation formula is as follows:

$$\beta = \mathrm{E}[(\frac{X-u}{\sigma})^4] \tag{4}$$

(3) Solve

Using Matlab to solve the descriptive statistical values for six categories, as shown in Table 3:

Table 3. Descriptive statistical analysis results of sales volume for vegetable categories

	minimum value	Maximum value	mean value	standard deviation	Skewness statistics	Kurtosis statistics
Aquatic rhizomes	0.926	296.792	37.402	31.357	2.508	12.162
Flower leaf class	31.29	1265.473	182.969	86.199	2.901	26.216
Chili peppers	6.066	604.231	84.413	53.436	3.146	18.420
edible fungi	3.012	511.136	70.126	48.489	2.996	16.896
florescent vegetables	0.632	186.155	38.511	22.672	1.529	4.208
Solanaceae	0.252	118.931	20.896	13.198	1.730	5.788

(4) Result analysis and conclusion

The standard deviation of aquatic rhizomes is relatively large, indicating that their data distribution is relatively scattered and skewed, indicating that their data distribution is significantly skewed to the right, and the kurtosis is also relatively large, indicating a high kurtosis of the distribution, and the data is relatively concentrated near the mean. The standard deviation of the flower leaf class is also large, with a positive skewness, indicating that its data distribution is also right skewed and has a high kurtosis, indicating a high sharpness of the distribution and a strong degree of data concentration [10]. The standard deviation of chili peppers and edible mushrooms is relatively large, but the skewness and kurtosis are larger than those of aquatic rhizomes and flowers, indicating that the data distribution of these two categories is more scattered and biased towards the right, with high kurtosis. The skewness is also high, indicating that the data in this category exhibits a clear positive skewed distribution. The standard deviation of cauliflower and eggplant is relatively small, indicating that their data concentration is relatively high, and the skewness and kurtosis are relatively small, indicating that these data are more evenly distributed than other categories, and there is no obvious skewness

2.7. Time Series

Time series analysis is a statistical method used to study the patterns and trends of data over time. It can help understand the periodicity, trend, and seasonality in time series data, and make future predictions through prediction and modeling.

(1) Sequence diagram

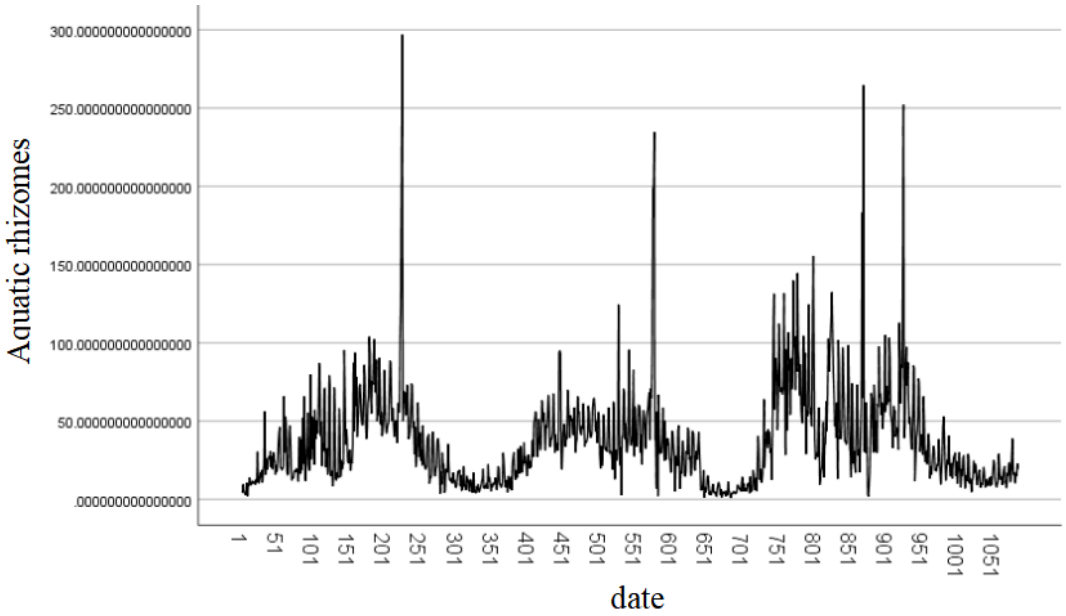

Figure 5. Sequence diagram of aquatic rhizomes

From Fig 5, the sales volume of each category is high in January and February each year, because the Spring Festival is usually in January and February, and people will shop and stock up during the Spring Festival. Therefore, analyzing the sales volume may have a certain relationship with holidays. The trend of aquatic rhizomes and edible fungi is similar, and based on the correlation coefficient between them, it is speculated that aquatic rhizomes and edible fungi are more commonly consumed in combination in daily life.

3. Conclusion

This article constructs a series of new data points to replace the independent variable values in the original data for each selected category by increasing or decreasing their values. Reduce the 3-year data to 1 year, analyze recent sales, observe the sensitivity of the dependent variable to changes in the value of the independent variable. According to the calculation results, there is no significant difference in the regression coefficient at the 95% confidence level, indicating that the model has good sensitivity. Time series analysis is susceptible to seasonal and other factors. We assign different weights to the data based on dates. The larger the weight of recent data, the more seasonal the data is decomposed. By adjusting the ratio of training data to validation data, combined with the factor analysis results of daily sales of different categories, we assign different weights to vegetable categories using factor scores. At the same time, when pricing, we consider the daily sales of different categories. The combined correlation coefficient of the total amount. By interpreting the results of factor analysis, one can gain a deeper understanding of the potential structure and correlation between different categories in vegetable sales data. These findings have guiding significance for developing marketing strategies, meeting consumer needs, and optimizing product portfolios.

References

[1] Wang Xiaotong Research on Inventory Control Strategy for Fresh Fruit and Vegetable Products in Chain Supermarkets Based on DIT [D] Southeast University, 2020.

[2] Zeng Minmin Research on Dynamic Pricing Strategy of Fresh Community Supermarkets Based on Time Scenario a [D] Southwest University of Finance and Economics, 2021.

[3] Chen Jun A Study on Ordering Strategy and Supply Demand Coordination of Fresh Agricultural Products Supply Chain Considering Circulation Loss Control [D] Chongqing University, 2009.

[4] Lin Tingting The vegetable sales model, problems, and countermeasures in Lishui District, Nanjing City [J] Shanghai Agricultural Technology, 2022 (01): 1 - 3.

[5] Zhang Fan Research on Agricultural Product Sales Strategies Considering Variable Opacity [D] Tianjin University, 2020.

[6] Cao Han Research on Predicting Sales of Catering Industry Dishes Based on Deep Learning [D] Xi'an University of Technology, 2019.

[7] Liang Shuting. A Study on the Community Fresh Retail Terminal Inventory Strategy with the Joint Influence of Time and Inventory Quantity on Deterioration Rate [D]. Beijing Jiaotong University, 2020.

[8] Mao Lisha. Research on Pricing Strategies and Production and Sales Models of Vegetable Wholesale Market from the Perspective of Supply Chain [D]. Central South University of Forestry and Technology, 2023.

[9] Zhou Xiangyu, Li Si. Short term prediction of express delivery business volume in Jiangsu Province based on TOPSIS criterion and SARIMA model [J]. Science and Industry, 2023, 23 (17): 136 - 142.

[10] Yang Yang, Tian Dingsheng, Zhang Bao'an, et al. Research on Urban Economy and Population Prediction Based on ARIMA Model [J]. Comprehensive Transportation, 2023,45 (11): 79-85+97.