

Replenishment Forecasting of Vegetable Commodities Based on Time Series Modeling

Baoluo Deng^{1, #}, Taotao Liu^{2, #} and Ruizhi Zhang^{3, *, #}

¹Faculty of Geography and Environmental Sciences, Guizhou Normal University, Guiyang, China

²School of Mathematical Sciences, Guizhou Normal University, Guiyang, China

³College of Teacher Education, Guizhou Normal University, Guiyang, China

*Corresponding author: 632812535@qq.com

#These authors contributed equally to this work

Abstract. The pricing and replenishment strategy of vegetable commodities affects the normal operation and economic returns of the superstore, if a reasonable automatic pricing and replenishment strategy system model of vegetable commodities can be developed, on the one hand, it can facilitate the merchants to carry out a reasonable management of the purchase and sale of vegetable commodities to maximize the economic benefits. On the other hand, merchants can provide a reasonable mix of vegetable products for customers to choose. In this paper, we first use the Apriori algorithm in the principle of correlation analysis to calculate the correlation relationship between different categories of commodities and single products under a certain confidence level. Then, we analyze the principle of cost-plus pricing method and derive the total sales and cost-plus pricing factors such as profit margin and cost relationship. Finally, a time series algorithm model is used to predict the total daily replenishment for the coming week.

Keywords: Correlation analysis; time series modeling; replenishment forecasting.

1. Introduction

Vegetables are one of the necessities of urban and rural residents, and commercial supermarkets are the main place for urban residents to obtain vegetable products, the supply and demand for its products and price changes deeply affect the pricing and replenishment decisions of the super. Generally speaking, the freshness period of vegetable commodities is short, and the quality deteriorates with the increase of sales time, and the products with shipping damage and poor quality are sold at a discount, and the products that are not sold on the same day cannot be sold on the second day. Although the variety of vegetables sold in the supermarket, the origin will be different, but the person in charge of the supermarket are not fully aware of the specific single product and the price of goods in the condition of the vegetable products on the day of replenishment. From the point of view of product demand, the sales volume of vegetable goods has a certain correlation with the time, from the point of view of product supply, April-October is a richer variety, but due to the limitations of the space of the superstore sales, it is necessary to carry out a reasonable combination of sales of products [1].

2. Correlation between different categories of vegetable products

2.1. Relevance analysis

Degree of support

Support (S) refers to the number of orders for both A,B as a percentage of the total number of orders. In business, it refers to how often and how many times a commodity or product is purchased by a customer in a certain period of time. It largely reflects the degree of trust and willingness of customers to buy the commodity. The formula for calculating the support degree is:

$$support(A \Rightarrow B) = P(A \cup B) \quad (1)$$

Where $P(A)$ is the percentage of orders in which the customer buys A and $P(B)$ is the percentage of orders in which the customer buys B. When a customer is purchasing a certain item, he or she may usually take into account the degree of approval of the item by other users. If the support level of a commodity is high, it means that more consumers are willing to buy and recognize the commodity, and consumers will think that the commodity is trustworthy, which increases the customers' desire to buy. On the contrary, if an item has a low level of support, consumers may be skeptical about it or choose to buy other similar alternatives [2].

Confidence level

Confidence refers to the probability of deriving B based on the association rule " $A \rightarrow B$ " if precondition A occurs. That is, it means the probability that Y will occur at the same time in the value of the term set where X occurs. The confidence level is calculated by the formula:

$$\text{confidence}(A \Rightarrow B) = P(B/A) \quad (2)$$

Where $P(A)$, $P(B)$ are the percentage of orders for customers to buy A and the percentage of orders for customers to buy B, respectively. When customers are purchasing goods, consumers usually compare different brands and models of goods, and will consider factors such as price, features, and quality. If the confidence level of an item is high, consumers may be more inclined to buy a certain type of item because they believe that this item can satisfy their needs, which affects their purchasing decision.

Promotion degree

Elevation is the ratio of the likelihood of event X occurring when event Y occurs to the likelihood of event B occurring when event A does not occur. In business, it is the ratio of "the proportion of orders for A that also buy B" to "the proportion of orders for B". The formula for the degree of lift is:

$$\text{Lift} = (P(A \& B) / P(A)) / P(B) = P(A \& B) / P(A) / P(B) \quad (3)$$

Where, A, B, respectively, represents two different commodities, $P(A)$, $P(B)$ for the customer to buy A order percentage, for the customer to buy B order percentage. Generally speaking, when the degree of enhancement >1 , it indicates that the customers who purchase commodity A are more likely to purchase commodity B than the overall average; when the degree of enhancement <1 , it indicates that the customers who purchase commodity A are less likely to purchase commodity B than the overall average; and when the degree of enhancement $=1$, it indicates that there is no correlation between the two different commodities of A and B [3].

2.2. Relevance Calculation Results

Table 1. Relevance Calculation Results

Relatedness	Brassica pekinensis	king cobra or ghost chili	shiitake, an edible mushroom	lettuce	Capsicum annuum	yellow cabbage
Brassica pekinensis		0.749	0.359	0.618	0.478	0.836
king cobra or ghost chili	0.666		0.168	0.429	0.874	0.847
shiitake, an edible mushroom	0.559	0.551		0.738	0.487	0.564
lettuce	0.830	0.891	0.636		0.735	0.384
Capsicum annuum	0.928	0.573	0.937	0.729		0.746
yellow cabbage	0.569	0.774	0.481	0.824	0.928	

As shown in Table 1, a correlation greater than 0.5 indicates that these items are related, and when a customer buys one of the vegetables, he/she buys the other vegetables in the correlation rule.

3. Relationship between total sales of vegetable types and cost-plus pricing

3.1. Modeling of correlation analysis

Since we are not sure about the relationship between the total sales volume of vegetable categories and cost-plus pricing, we decided to describe and analyze the correlation between the total sales volume of vegetable categories and cost-plus pricing, and to determine whether they satisfy the conditions of the correlation analysis, so as to determine whether there is a relationship between them.

(1) Whether both variables are quantitative data

Quantitative data refers to data that can be expressed in terms of quantity, amount or other measurable units. In terms of total sales, we can calculate the number of vegetable products sold to arrive at a specific figure, for example, 10 kilograms of vegetables sold. And in terms of cost plus pricing, we can similarly transform the cost and pricing of a vegetable product into specific numbers, for example, the cost of a vegetable is \$12 per kilogram and the pricing is \$18 per kilogram. Usually these data are quantifiable, so both total sales and cost-plus pricing of the vegetable category are quantitative data.

(2) No outliers between the two variables

We used SPSS software to find outliers in the total sales volume and cost-plus pricing of vegetable categories, and found that there are no outliers between the two variables.

(3) There is a linear relationship between the two variables

The use of scatterplots for testing is often one of the most important means of determining whether there is a linear relationship between two variables, and we imported the data of the total sales volume and cost-plus pricing of vegetable categories into Excel software to obtain the following scatterplot:

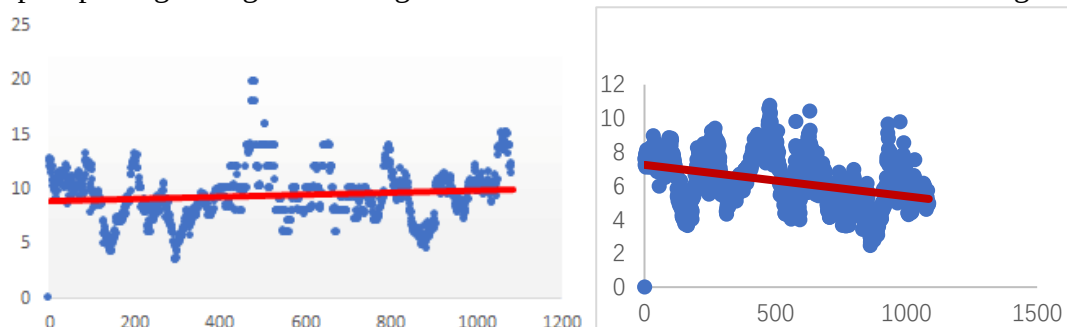


Fig. 2 Cauliflower sales distribution **Fig. 3** Distribution of sales volume of flowers and foliage

As shown in Figures 2 and 3, the scatterplot shows a relatively clear linear relationship between total sales volume and cost-plus pricing for different vegetable categories, with cost-plus pricing for vegetables roughly leveling off as sales volume increases.

(4) Whether the two variables conform to normal distribution

Normal distribution is a method of inference based on linear relationship, which can ensure that the mean and variance of the sample conform to its characteristics, thus making the results of inference more stable and reliable. We imported the data of total sales and cost-plus pricing of different vegetable categories into SPSS software, and the effect of the normality test that we obtained.

Because the sample size is less than 50, we use the results of the K-S normality test in SPSS, specifically, the total sales of five types of vegetable categories such as cauliflower, eggplant and other categories of vegetables with cost-plus pricing corresponds to a p-value of less than 0.05, which does not show significance, which indicates that it accepts the original hypothesis that the two variables obey a normal distribution.

3.2. Pearson correlation analysis

Since there are five categories of vegetable categories whose total sales and cost-plus pricing satisfy the Pearson correlation analysis, we use the Pearson correlation analysis to calculate the

correlation coefficients between total sales and cost-plus pricing for the vegetable categories using the formula:

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}} \quad (4)$$

Where r represents the Pearson's correlation coefficient of the two variables. $\sum$ represents the summation of all the data. X and Y represent the observed values of the two variables. $\bar{X}$ and $\bar{Y}$ represent the mean of the observed values of the two variables respectively. A and A represent the mean of the observed values of the two variables [4].

After calculating the correlation coefficient $r=1$ significance test, if the Pearson's correlation coefficient is equal to 1, this means that there is a strong positive linear relationship between the two variables. In other words, when the sales volume increases, the price also increases at the same rate.

3.3. Spearman correlation analysis

We found that although the eggplant data in the vegetable category is not normally distributed, so we used Spearman correlation analysis to discriminate the relationship that exists between sales volume and price in the eggplant data. The formula for this is:

$$p = 1 - (6 \cdot \sum d^2) / (n \cdot (n^2 - 1)) \quad (5)$$

Where P denotes the Spearman correlation coefficient, $\sum d^2$ denotes the sum of the squares of the differences of all corresponding rankings, and n denotes the sample size.

Substituting the formula and calculating $P=1$ indicates a strong positive correlation between the two variables, i.e., as the sales volume of eggplant increases, the sales price variable of eggplant increases as well, and the magnitude of the increase remains relatively consistent.

4. Relationship between total sales of vegetable types and cost-plus pricing

We will use time series analysis to forecast the total daily replenishment for the coming week based on this data on sales of six different vegetable categories from July 2020 - July 2023, using time series analysis.

4.1. Descriptive statistical analysis

We calculated the mean, standard deviation, maximum and minimum values of six different vegetable categories through SPSS software, and found that the trend of change in the sales volume of cauliflower, cauliflower, chili pepper, eggplant and edible fungus vegetables showed the characteristics of seasonality, cyclicity and trend, so we can choose the appropriate time-series model for forecasting and planning of sales strategy.

4.2. Visual analysis

As shown in Figure 4, we use SPSS software to draw a line graph of the time series data to observe the trend of the data of the six different vegetable categories, and through the analysis, it is concluded that the sales volume of the seed different vegetable categories has the characteristics of seasonality and cyclicity.

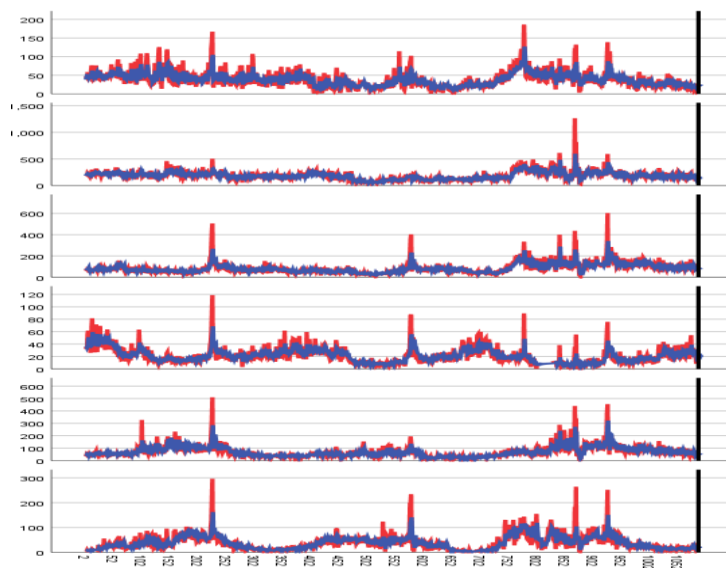


Fig. 4 Line Chart of Different Vegetable Categories

As can be seen from the figure, this time series has a more obvious trend and does not contain irregular fluctuation series with periodic changes, i.e., it is a non-stationary time series, therefore, we forecast and analyze this time series with the non-stationary time series analysis method.

The model is adopted using the model: $X_i = \mu_i + Y_i$, where μ_i denotes the value of X_i that changes over time, and Y_i is the remaining part of X_i after excluding μ_i . Here we know through analysis that Y_i can be ignored.

From the figure, it can be seen that there is no obvious regularity in the change of the data, showing a trend of randomness, therefore, we choose the following methods to solve the problem.

(1) Make residuals. According to the value of the fitted μ , find the residual sequence, the residual sequence graph is shown in Fig. 3. Observing the residual series in Fig. 3, we can see that the series has large volatility, so it can be considered non-smooth, and can be smoothed by doing multiple differencing.

(2) Make difference. The residual series $Y_a = X_a (a=1,2,...,26)$ is differenced to smooth it.

(3) The expert model in SPSS software is used to describe the relationship between the variables based on existing expertise and experience, and automatically find the optimal model order, thus determining that it is the AMRA (1, 2) model.

(4) It estimates the parameters in the probability distribution by selecting the parameter values that maximize the probability of occurrence of the observed data. In SPSS, maximum likelihood estimation can be performed using statistical analysis tools, e.g., by selecting the appropriate statistical model, setting up the variables and observations, and then running the appropriate analysis commands to perform maximum likelihood estimation [5].

4.3. Testing of the model

In SPSS, we use linear regression analysis to calculate the residuals and perform a test of the residual vector to obtain a test value of 0.003 for the residual vector, which is roughly close to 0, indicating that the model is usable. The total daily replenishment for each vegetable category for the coming week can also be derived as shown in Table 2 below.

Table 2. Total daily replenishment for each vegetable category for the coming week

Number of days	1	2	3	4	5	6	7
Cauliflower	23.42	21.89	20.88	21.95	22.95	23.68	24.83
Philodendron	132.49	126.54	126.64	139.88	148.14	140.91	142.94
Capsicum	79.62	80.21	80.50	82.16	81.81	84.43	86.32
Eggplant	22.15	22.10	20.19	21.07	19.54	21.15	19.95
Food category	43.14	46.18	49.58	53.97	53.39	52.73	50.83
Aquatic rhizomes	21.43	20.00	20.95	20.62	20.99	20.97	21.16

5. Summary

In this paper, the Apriori algorithm is used to mine the possible existence of certain associative relationships between different categories or different single products of vegetable goods, and its idea is simple and easy to understand, expandable, efficient in processing data, and interpretable. However, we randomly choose one day to analyze the possible correlation relationship between different categories or different single products of vegetable category commodities, and the result of the analysis has a certain degree of chance.

In this paper, Pearson correlation analysis was used in analyzing the relationship between sales volume of vegetable categories and cost-plus pricing. However, Pearson correlation analysis can only detect linear relationships, but correlation does not imply causation. Other potential factors, such as seasonality, consumer habits, etc., need to be considered when examining the relationship between the total sales volume of each vegetable category.

Finally, we predicted the sales volume for the coming week using a time series algorithm.

References

- [1] Shen, Dongjie. Collaborative management of inventory and replenishment strategies in perishable products supply chain[D]. Anhui:University of Science and Technology of China,2009.
- [2] Pan Zhang, Lingyun Ding, Ning Jiang, et al. Data mining algorithm and application of distribution network automation system based on support-confidence-enhancement[J]. Electrical Measurement and Instrumentation,2019,56(10):62-68.
- [3] GUO Yanping,GAO Yun,JING Wen. Research on correlation analysis of air pollutants based on Apriori algorithm[J]. Software Engineering,2023,26(09):8-11+32.
- [4] ZHAO Ziqi, QU Ningsong. Correlation study of microgroove machining parameters and groove width in electrolytic milling of tube electrodes[J]. Aerospace Manufacturing Technology,2021,64(9):39-44.
- [5] YANG Jingyun, LIAN Feihao, LU Jiye et al. A dynamic production optimization model based on time series prediction[J]. Mathematical Modeling and Its Applications,2023,12(02):115-124.