Information Systems and Economics (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/infse.2024.050303
ISSN 2523-6407 Vol. 5 Num. 3

Analysis of Vegetable Pricing and Replenishment Strategies Based on ARIMA Time Series and Random Forest Algorithm

Junxiao Chen[1,*], Kaize Wang[1], Yufan Li[1]

[1]*School of International, Jilin University of Finance and Economics, Changchun, China*
[*]*Corresponding author: 1923975983@qq.com*

Keywords: Pricing and Replenishment Strategy, ARIMA Time Series, Random Forest Algorithm

Abstract: Fresh food superstores need to take advantage of the freshness period of vegetables and the corresponding time of day to purchase, using pricing methods and replenishment strategies based on the cost-plus pricing method. Proper pricing and replenishment strategies can maximize the supermarket's revenue. In this paper, we analyze the data of different categories of vegetables and different single products to get their distribution patterns and correlations, use ARIMA time series to predict the future wholesale price, use linear regression to predict the future demand, and then based on the Random Forest algorithm to predict the future sales volume of each category and single product, and then the final pricing of the product and the replenishment strategy is based on the premise of maximizing the revenue, and based on the optimization model.

1. Introduction

In life, superstores are one of the main players in the sale of fresh vegetables, however, the management problems for vegetable products pose considerable challenges. These challenges stem from the fact that fresh vegetables are perishable and highly influenced by the seasons. Supermarkets need to forecast the sales trend of vegetables [1], set reasonable prices based on cost-plus pricing and other strategies, and introduce an optimized product mix and replenishment strategy that considers seasonal variations.

In this paper, we first analyze the distribution patterns and interrelationships of vegetables and sales volume of different categories and single items. Using the pricing formula [2], the cost profit margin of each single item is obtained. Then using the weighting method to assign the importance of each single product relative to the category to which it belongs combined with its sales volume, the relationship between sales volume and cost margin in the six categories is obtained. The relationship between price and sales volume was obtained by running the random forest algorithm. Under the condition that the future price trends of the six categories are predicted by the ARIMA time series [3], the optimization model is established by subtracting the wastage caused by the wastage rate from the profit. The planning model is utilized with the constraints of the total number of saleable items and the minimum number of columns. At the same time, the wastage rate is considered to find a

profit-maximizing pricing and replenishment strategy that can be implemented by the superstore under the condition of satisfying the demand for the six categories of vegetables as much as possible.

2. The distribution pattern of the category and the correlation between single products

In this chapter, the acquired data is visualized into the basic situation of categories and individual items, the overall distribution pattern of categories and individual items, and the time pattern of categories by means of histograms and line graphs, etc., which in turn allows us to organize the information about the flow of individual items into different orders through the change of time, and run the Apriori algorithm for the discovery of association rules.

2.1 Basic information on categories and single items

There is some variation in the category distribution of individual items in different categories, and in order to measure this variation, this paper visualizes the number of individual items in different categories in a histogram, as shown in Figure.1. So, it is seen that the largest variety is foliage, followed by edible fungus, followed by chili, aquatic rhizomes, eggplant, and cabbage in that order.

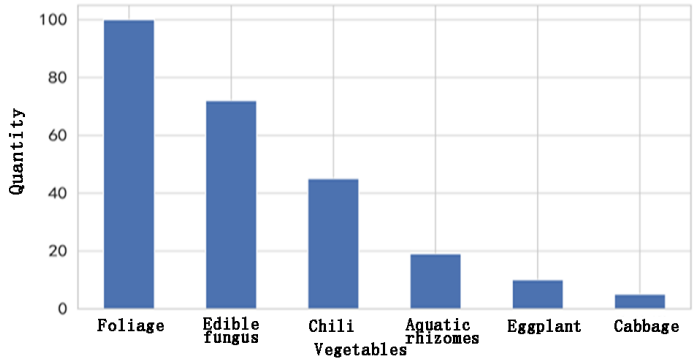

Figure 1: Distribution of number of types in different categories

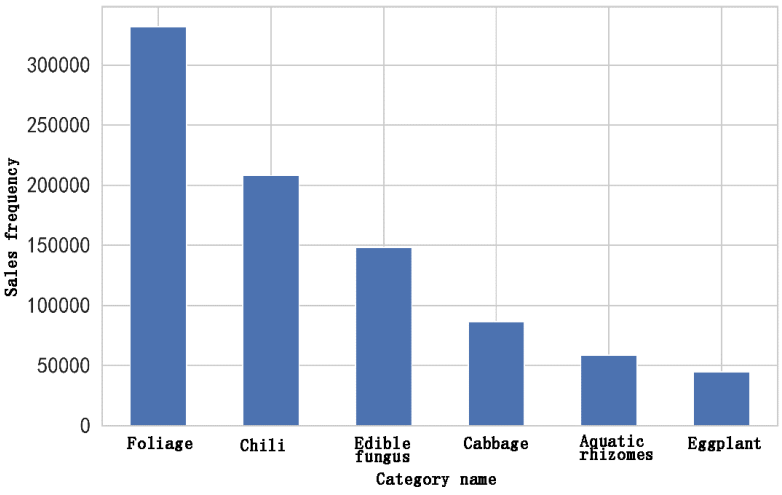

Figure 2: Category sales frequency

Although there are many types of foliar categories, the actual purchase, may not be the more types of vegetables will be purchased more frequently, therefore, this paper counts the number of purchases

of different categories of vegetables in the data, by the principle of appearing once to increase once for the visualization of the histogram, the specific situation is shown in Figure 2.

Further, we counted the purchase frequency of all individual items to see exactly which items made the main contribution to the purchase frequency of the category, as shown in Figure 3.

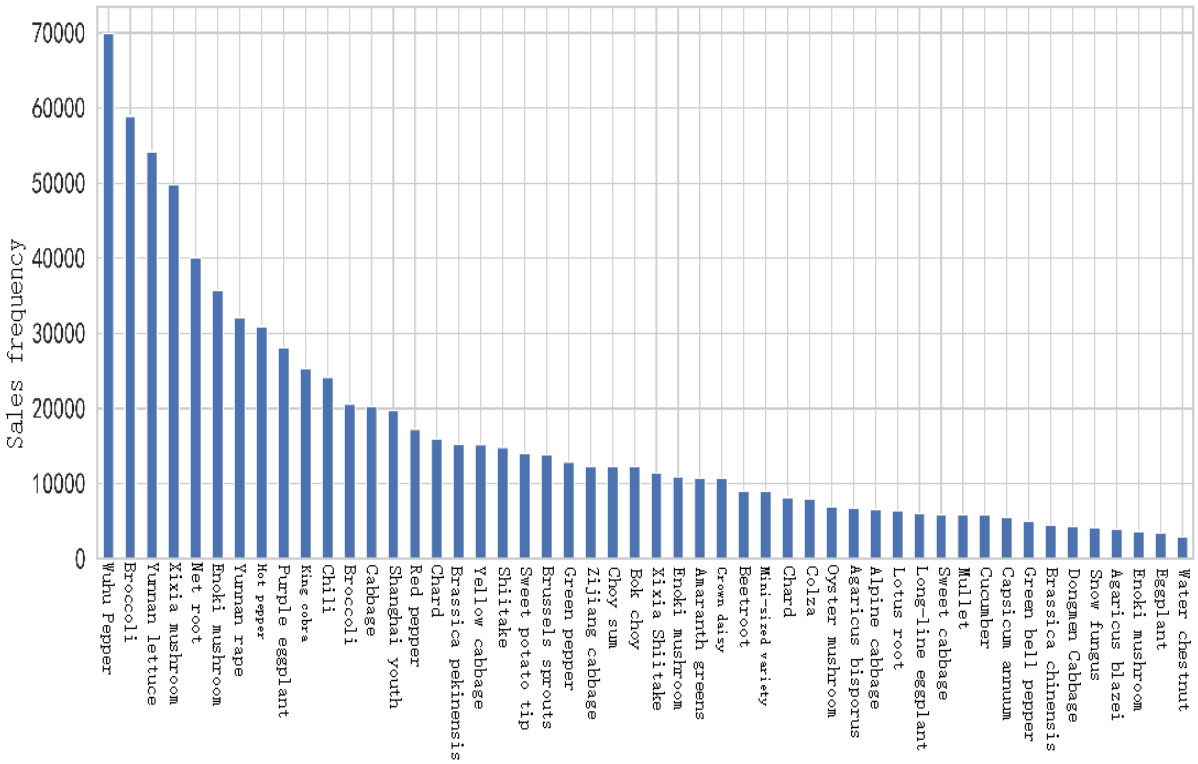

Figure 3: Single item sales frequency

It can be seen that the most frequently purchased item is Wuhu pepper in the chili category, and the second is broccoli in the foliage category. Enoki mushroom in the aquatic root category then also appeared at the top, but overall, most of them were singles in the chili and foliage categories.

2.2 The distribution pattern of sales time of the different category

In addition to understanding the overall sales of different categories and single items, this section also analyzes the sales of different categories at different time periods from the time dimension, which helps to develop replenishment strategies. For example, for products with high sales in the morning, replenishment can be considered in the early morning; for products with high sales in the afternoon, replenishment can be carried out at noon. Considering the large number of single items in different categories and the fact that a category is a collection of single items, we chose to visualize the sales at different times at the category level and the results are shown in Figure 4.

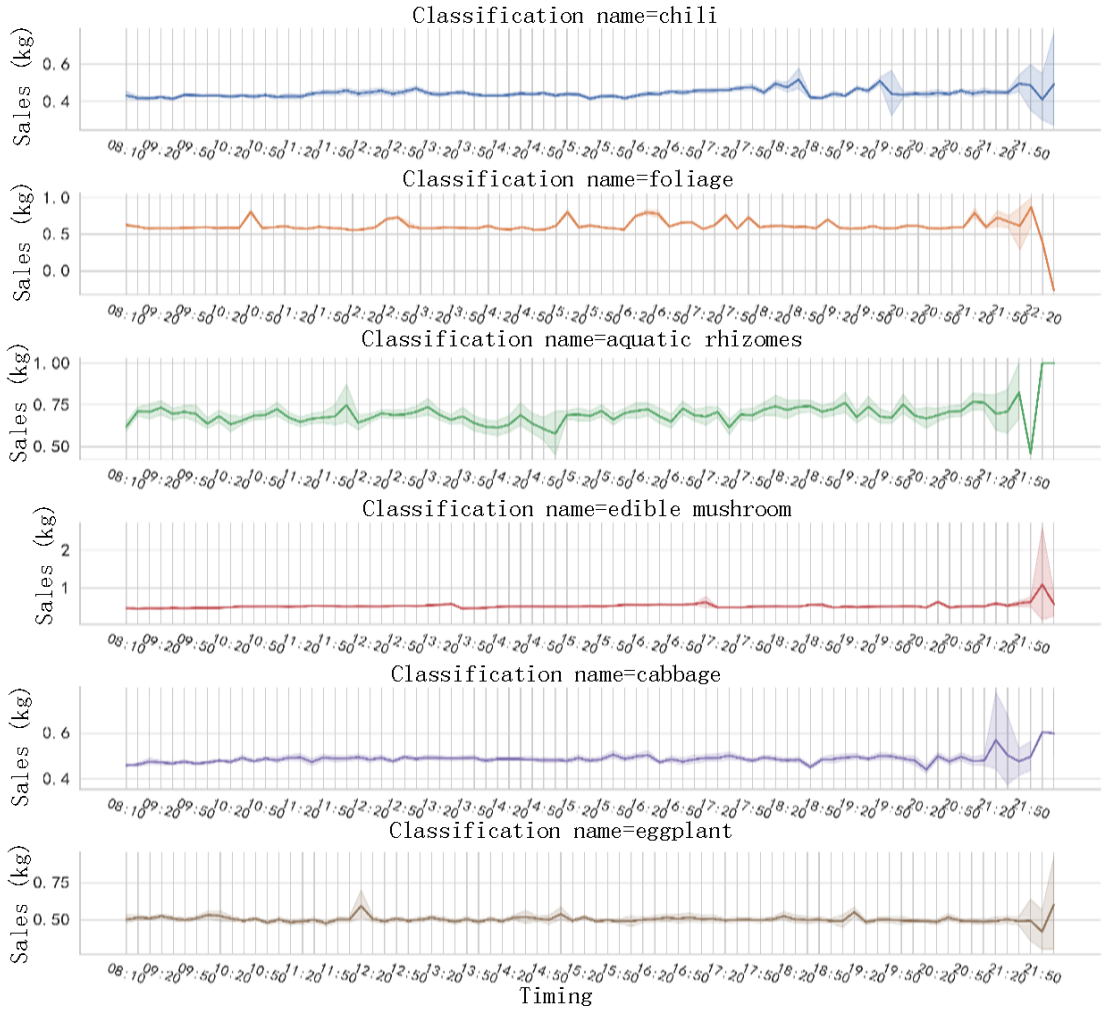

Figure 4: Sales distribution of different categories at different time periods

The above figure demonstrates that there are significant differences in the sales volume of different categories at different time periods. A common feature is the large fluctuation of data at the end of the period, which is mainly since when the supermarket is about to close, the number of customers entering the store to buy vegetables is very small, leading to a significant reduction in the amount of data, which in turn causes large fluctuations.

For chili products, the sales trend is relatively smooth, but there are small fluctuations at 6:00 pm and 7:00 pm and 8:00 pm. This time period is after mealtime, and it is speculated that some merchants who sell at night may buy these products and therefore purchase chili products in the evening.

There were four main fluctuations in the foliage category, at 10:50, 1:00 pm, 4:00 pm, and 6:00 pm. These time periods were either close to or after mealtime, which is a normal fluctuation in sales.

The fluctuations for aquatic rhizomes products were more specific. We believe this may be due to the strong seasonal factor and less stable supply of aquatic root products.

Edible mushrooms, cabbage and eggplant products were purchased at a more even time, which is a normal sales situation.

2.3 Association rules for single items in the same category

Individual items purchased near each other are included in one order and belong to the same shopping basket, as inferred by the time of payment. We merge the neighboring single items in ten

second intervals into the same order. Then using Apriori algorithm, association rule mining is performed.

The two key parameters of Apriori algorithm principle are minimum support and confidence [4], the minimum support helps to help us quickly eliminate the item sets smaller than the minimum support, and the confidence helps us to get the final strong association rules. We set the minimum support to 50 and the confidence level to 60.

We solved the problem in Python and found that there is no strong association rule with a confidence level of 0.6. We think the reason is that the screening result is to satisfy the confidence requirement. In this paper, we still use the Apriori algorithm to filter the frequent item set for correlation analysis, we were based on the support of the frequent item set of the top four single items, three single items and two single items combination of single items shopping basket.

According to the results, it can be known that Yunnan lettuce, broccoli, Xixia shiitake mushrooms and net lotus root appear in most of the values of the most frequent item sets in the four item combinations, and the highest ranking is Yunnan lettuce, broccoli, Wuhu green pepper and net lotus root, followed by broccoli, Wuhu green pepper, Xixia shiitake mushrooms and net lotus root. The strongest correlation is between broccoli and Wuhu green pepper. The set of highly supported frequent items for the three single-item combinations is relatively consistent with the results for the four single-item combinations, with the only difference being the occurrence of the pairing of purple eggplant with broccoli and Wuhu green pepper, and the pairing of Yunnan oleander with Wuhu green pepper and broccoli. The ordering of the frequent item sets for the two single-item combinations shows that the highest degree of co-occurrence is with western blue flowers and Wuhu green peppers, followed by Yunnan lettuce and western blue flowers, and the combinations with higher frequency are also with western gorgonzola mushrooms, nettles, and purple eggplants.

3. Pricing and replenishment strategies

The cost margin for each individual product will be found by the cost-plus pricing formula, and secondly the relationship between sales of individual products and cost-plus pricing will be informed by the Spearman test. In everyday life, the sales volume of a product with low sales volume has a similarly weaker effect on pricing and replenishment. This qualifies the application of the weighting method. Therefore, the wholesale price, sales price and cost margin of different individual products will be weighted. Finally, the relationship between sales and cost-plus pricing for the different categories will still be determined by the Spearman test.

Price affects consumers' willingness to buy, which in turn affects sales, and supermarkets' profits are affected by sales. This leads to the importance of developing a mathematical model of the relationship between price and its sales volume for different categories of vegetables. At the same time, the wastage rate is also one of the influencing factors. Taking the above conditions into consideration, the problem of formulating pricing and replenishment strategies that enable the superstore to capture the maximum profit will be given a solution by the optimization model.

3.1 Relationship between total sales volume of different categories of vegetables and their cost-plus pricing

The cost margin for each individual item will be found by the cost-plus pricing formula:

$$rate_a^b = (sp_a^b - w_a^b)/w_a^b \tag{1}$$

In this formula, $rate_a^b$ represents the cost margin of item a in category b, sp_a^b represents the selling price of item a in category b, and w_a^b represents the wholesale price of item a in category b.

We aggregate the records of the same item purchased at different times and sum the sales volume to get the total sales volume t_a^b and then we calculate the total profit $profit_a^b$. The formula is as follows.

$$profit_a^b = t_a^b \times (sp_a^b - w_a^b) \tag{2}$$

The four components of the formula are affected by time, which varies the results obtained. The question is in the context of an unknown data distribution and requires us to analyze the relationship between the cost margin of a single product and its sales volume. Spearman's correlation coefficient can solve this problem. For category correlation analysis, we aggregate individual items in the first step and sum their total sales and total profit in the second step. The formula is as follows.

$$t^b = \sum_{a \in b} t_a^b \tag{3}$$

$$profit^b = \sum_{a \in b} profit_a^b \tag{4}$$

Next, the normalization coefficient coe_a^b will be calculated for item a in category b, and the result will be obtained from the percentage of sales. The formula is as follows.

$$coe_a^b = t_a^b / \sum_{a \in b} t_a^b \tag{5}$$

Next calculate w^b, sp^b, $rate^b$ as shown below:

$$w^b = \sum_{a \in b} coe_a^b \times w_a^b \tag{6}$$

$$sp^b = \sum_{a \in b} coe_a^b \times sp_a^b \tag{7}$$

$$rate^b = \sum_{a \in b} coe_a^b \times rate_a^b \tag{8}$$

The different individual products and categories were analyzed using Pearson correlation analysis.

It is concluded that there is not a strong correlation between cost margin and total sales for both individual items and categories. However, in the heat map where the categories are cauliflower and chili, we can see that the total profit becomes larger as the cost margin increases. But in our economic life, when the increase of cost margin means the increase of price, when the price is high to a certain level, the sales volume will decrease, and the total profit will also decrease. We can get from reviewing the literature that vegetable products belong to the necessities of life, which are less elastic and less affected by the price. Therefore, an appropriate increase in cost margin can increase the total profit of the superstore.

3.2 Pricing and replenishment strategies for different categories for the coming week

First, we use the ARIMA time series model [5] to predict the price trend in the coming week based on the data related to wholesale prices. Second, we determine the order of the ARIMA model, which can be found by searching the internet and setting the search interval to get the MA and AR orders. Only the cauliflower category is shown in Figure 5.

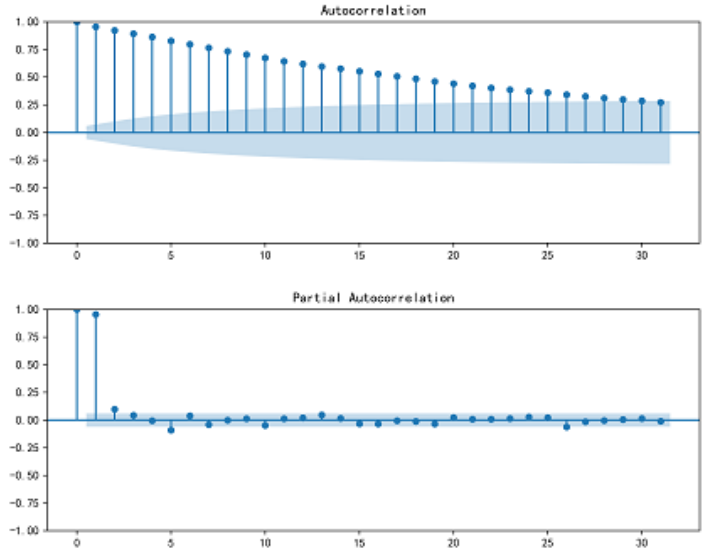

Figure 5: ARIMA time series of cauliflower category

To obtain more accurate results, we turned our attention to the random forest model [6]. Firstly, we combine the four data with temporal features together, secondly, training set and testing machine are divided in the category data, then testing and training are performed and finally training set and testing results are produced.

The random forest model allows us to calculate the total profit with a fixed selling price. We find that it can be solved with the optimization model. The formula is as follows.

$$Max \ \ \hat{t}^b \times (\widehat{sp}^b - \hat{w}^b) \tag{9}$$

$$s.t \ \ \hat{w}^b \leq \widehat{sp}^b \leq 3\hat{w}^b \tag{10}$$

$$\hat{e}^b \times \hat{t}^b \leq f^b \tag{11}$$

Where $\hat{t}^b$ represents the random forest predicted sales volume, $\widehat{sp}^b$ represents the pricing methodology, $\hat{w}^b$represents the wholesale price predicted by the ARIMA model, $\hat{e}^b$ represents the wastage rate of vegetable b products, f^b represents the replenishment strategy.

Finding the profit maximizing pricing and replenishment strategy is a linear optimization problem and we take a linear optimization approach to solve it.

4. Conclusions

Using a combination of mathematical and statistical techniques, superstores can achieve greater excellence in vegetable merchandise management, providing a powerful tool for maximizing profit growth and market competitiveness. First, time series analysis is one of the keys. By applying time series models such as ARIMA, various trends and seasonality in historical sales data are deeply explored to more accurately forecast future demand and price fluctuations. Second, regression analysis provides the ability to understand the factors influencing sales. Through models such as linear regression, identify the factors that have the greatest impact on sales volume, such as price, promotions, and advertising expenditures. Pricing strategies and promotional decisions are continually optimized to maximize sales and profits. Random forests and machine learning techniques,

on the other hand, introduce more advanced analytics that deal with relevant and complex multidimensional data, including supply chain efficiency, customer preferences, and market trends. These algorithms are utilized to discover non-linear relationships and hidden patterns to better predict sales volume and product mix success. Finally, optimization models play a key role. With the goal of maximizing profits, methods such as linear optimization are used to develop optimal product mix, replenishment strategies, and price pricing strategies.

References

[1] Mclaughlin E W. The dynamics of fresh fruit and vegetable pricing in the supermarket channel [J]. Preventive Medicine, 2004, 39(supp-S2): 81-87. DOI: 10.1016/j.ypmed.2003.12.026.

[2] Zheng A, Fang Q, Zhu Y, et al. An application of ARIMA model for predicting total health expenditure in China from 1978-2022 [J]. Journal of Global Health, 2020, 10(1). DOI:10.7189/jogh.10.010803.

[3] Cicchetti C J, Foell W K. Energy Systems Forecasting, Planning and Pricing [J]. Nasa Sti/recon Technical Report N, 1975.

[4] Hegland M. The apriori algorithm–a tutorial[J]. Mathematics and computation in imaging science and information processing, 2007: 209-262.

[5] BV B P, Dakshayini M. Performance analysis of the regression and time series predictive models using parallel implementation for agricultural data[J]. Procedia Computer Science, 2018, 132: 198-207.

[6] Bćhir Bęaoui, Armi Z, Ottaviani E, et al. Random Forest model and TRIX used in combination to assess and diagnose the trophic status of Bizerte Lagoon, southern Mediterranean [J]. Ecological Indicators, 2016. DOI: 10.1016/j.ecolind.2016.07.010.