DOI: 10.23977/infse.2024.050202
ISSN 2523-6407 Vol. 5 Num. 2

Fresh Vegetable Sales and Pricing Forecasting Based on Systematic Clustering and ARMA Modeling

Yichen Du

School of Mathematical Sciences, Henan Institute of Science and Technology, Xinxiang, 453003, China

Keywords: Hierarchical Cluster Analysis, ARMA model, autoregressive moving average model, sales forecast

Abstract: In this paper, in order to explore the distribution pattern and interrelationship among fresh vegetables and to make replenishment decisions for each vegetable category on the same day without exactly knowing the specific single product and purchase price, we use the systematic clustering model and the ARMA model to derive the similarity degree of the sales situation of different single products of vegetables, and the replenishment quantity and pricing strategy for the next seven days of the different categories of sales volume and cost-plus pricing, and we verify the reasonableness of the model by using the leafy and flowery vegetables as an example. The rationality of the model was verified with the example of leafy vegetables, yielding forecast relative errors of 0.1934 and 0.3334, which are in line with the expected situation. From the demand side, the study enables consumers to buy fresh vegetables on the same day. From the supply side, supermarkets can reduce unnecessary waste due to the perishability of vegetables.

1. Introduction

Fresh produce, including fruits and vegetables, plays a vital role in our daily life and the food industry. As people's living standards improve, the demand for fresh produce has also increased. This has led to an increase in the business position of supermarket fresh goods. However, since the supermarket is unaware of the specific single product and the purchase price, and due to the perishability of the vegetable category, the goods can become stagnant and require discounting or discarding. This results in unnecessary waste and economic loss. Furthermore, consumers are usually sensitive to freshness, and their purchasing decisions are based on price and freshness. Therefore, it is crucial to develop a scientific replenishment and pricing strategy based on historical sales and demand for each product.

Manish Shukla develops an autoregressive integrated moving average (ARIMA) model to forecast the demand for fresh produce daily and validates the model using sales data from the Indian onion market [1]. Girish K.Jha -Kanchan Sinha provides temporal forecasting of wholesale prices of oilseed crops in India using forward routed time-delay neural network (TDNN) and ARIMA models, comparing the merits and demerits of the two models in the linear case and the nonlinear case [2]. Dharavath Ramesh et al. used seasonal ARIMA to forecast the prices of fruits and vegetables, and formulated an appropriate strategy to reduce the prices of fruits and vegetables by using the forecasting results [3]. Tijun Fan and others solved a dynamic programming model, proposed a dynamic pricing strategy for multiple batches of fresh produce formulated a replenishment policy, and utilized four heuristic replenishment strategies to simplify replenishment operations [4]. Luyao Wang et al. summarize temporal forecasting models for agricultural commodities and suggest that the application of forecasting models based on price-influencing factors should be further expanded [5]. Helin Yin et al. proposed STL-ATTLSTM (STL-ATTLSTM, Attention-based LSTM) model integrating seasonal trend decomposition using loess (STL) preprocessing method and attention mechanism based on long and short-term memory (LSTM) and applied it to five crops, cabbage, radish, onion, pepper, and garlic, to compare the performance of the model with other models to comprehensively compare the performance advantages and disadvantages [6]. Tae-Woong Yoo and Il-Seok Oh used Seasonal Long Short-Term Memory (SLSTM) to predict weekly, monthly, and quarterly crop history data for forecasting [7]. Sourav Kumar Purohit et al. use two additive and five multiplicative hybrid methods to forecast tomatoes in India, onion, and potato monthly retail and wholesale prices and synthesize the comparison to derive the advantages and disadvantages of the forecasting results [8].

In this paper, to ensure the freshness of fresh vegetables in supermarkets, we summarize the laws and interrelationships between different vegetable categories, improve the shortcomings of the previous literature that only considers the prediction and pricing of single vegetables, establish an autoregressive moving average (ARMA) model to predict the total daily replenishment of each category and a vegetable pricing strategy based on cost-plus pricing, which compensates for the problem that the type of pricing in the previous research is not clear. Finally, the article takes the leafy vegetables as an example to verify the rationality of the model and the feasibility of the algorithm.

2. Model construction

2.1 Hierarchical Cluster Analysis

Clustering is an unsupervised learning method, and the basic idea of Hierarchical Cluster Analysis is to cluster the variables that are close to each other into classes according to their distance, and then cluster the variables that are farther away from each other, and gradually carry out the clustering process until each variable is categorized into a suitable class. Therefore, the basic idea of spectral cluster analysis of different vegetable individual products is to cluster systematically from multiple samples in a batch of vegetable samples, and to select the total sales of different individual products as the P meta-observations from among the processed data, when the conditions should be realized：

$$d(x_i,x_j)\geq 0, \tag{1}$$

$$d(x_i,x_j)=d(x_j,x_i), \tag{2}$$

$$d(x_i,x_j)\leq d(x_i,x_k)+d(x_k,x_j), \tag{3}$$

where x_i implies a point in the space of P elements, which is the sales volume of each single vegetable item, $d(x_i,x_j)$ is the distance between the sample points of the calculation. If $d(x_i,x_j)=0$ holds then if and only if $x_i=x_j$ holds.

In this paper, Euclidean squared distance is used as a distance measure for two different vegetable singles:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{P} (x_{ik} - x_{jk})^2}, \tag{4}$$

According to the principles of cluster analysis, to categorize the sales volume of different vegetable individual products and ensure the accuracy of clustering, the article observes the following rules for establishing systematic clustering:

(1) Maximize intra-class similarity: when clustering the sales volume of each vegetable item, the similarity coefficient is used to measure the degree of similarity between the sales volumes of different vegetable items. When the similarity coefficient is larger, the degree of similarity between each class is higher.

(2) Minimize interclass similarity: the average of the distances between all the samples in the two classes is used as the distance between the two classes, i.e., the class mean distance is used as the interclass distance to measure the similarity between the classes, and the similarity between the classes is smaller when the class mean distance is larger.

2.2 ARMA time prediction model

The ARMA model, known as the autoregressive moving average model, combines the autoregressive model (AR) and the moving average model (MA). The basic idea of the ARMA model is to characterize a time series by both autoregressive and moving average aspects, i.e., the current value of the time series is related to its past values and its past errors while maintaining smoothness. The following equation defines the ARMA (p, q) model:

$$x_t = \phi_0 + \phi_1 x_{t-1} + \cdots \phi_p x_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q}, \tag{5}$$

where x_t represents the value of the time series x at time t, $\phi_1, \ldots, \phi_p$ is the autoregressive coefficient, $\theta_1, \ldots, \theta_q$ is the moving average coefficient, and ε_t is a sequence of independently and identically distributed random variables.

The ARMA model has four main basic steps as follows, the specific steps are shown in Figure 1:

(1) The original time series is tested for smoothness, if the time series is smooth then go to (2), or else the time series needs to be differenced.

(2) Based on the correlation characteristics, the model type is identified using the truncated tails of the autocorrelation function (ACF) and the partial correlation function (PACF), and the lag order is used to order the model.

(3) Estimation of model parameters to produce the desired predictive model.

(4) Predictions were made using the established model and RMSE flat mushroom predictions were applied.

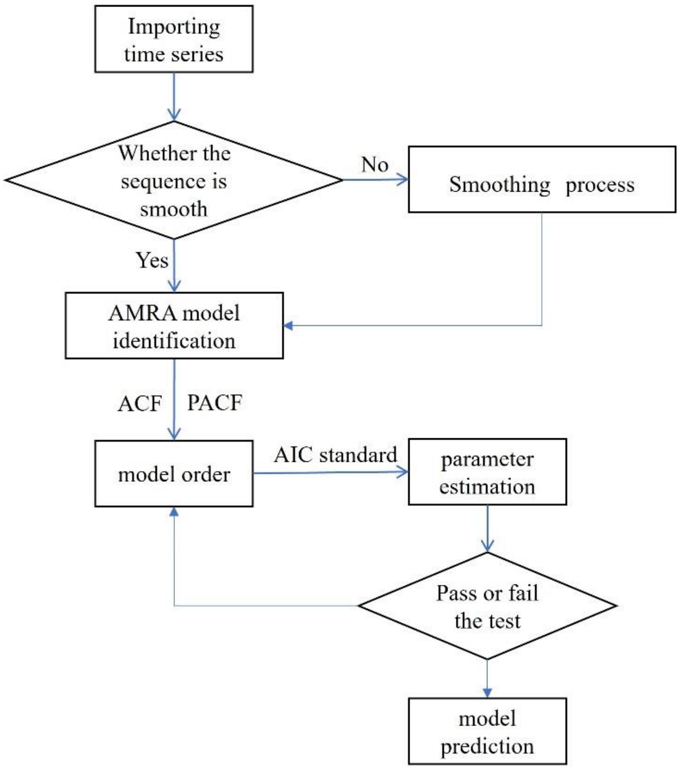

Figure 1: ARMA Model Flowchart

3. Results

3.1 Hierarchical Cluster Analysis Results

In this part, through the above clustering modeling, the clustering results are obtained as shown in Figure 2. It was found that there was a strong correlation between fresh rice dumpling leaves (bag), lotus root, and fruit chili (orange); and between amaranth, zhi jiang red beet moss (portion), lotus root tip, and golden needle mushroom (bag). The similarity between these two clusters was low. The following results suggest that people may tend to buy similar vegetables in one piece when purchasing fresh fruits and vegetables, and this result can be referred to when restocking and pricing for sale.

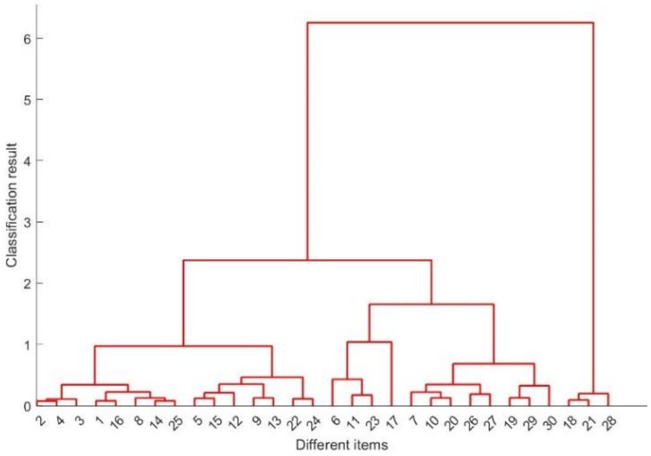

Figure 2: Hierarchical Clustering Image Results

3.2 ARMA Model Results

In this part, the research uses leafy and flowering vegetables as an example for time series forecasting of sales volume and price based on cost-plus pricing. First of all, for sales of flowers and foliage, the original data is processed and the autocorrelation function and partial correlation function are calculated, it can be seen that the original data of sales volume in Figure 3 is an unstable series, which changes significantly over time, and after differential reduction, a smooth series is obtained as shown in the right panel of Figure 3. Although there is still some fluctuation in the time series in the right panel of Figure 3, the change in the overall trend of increasing and then decreasing.

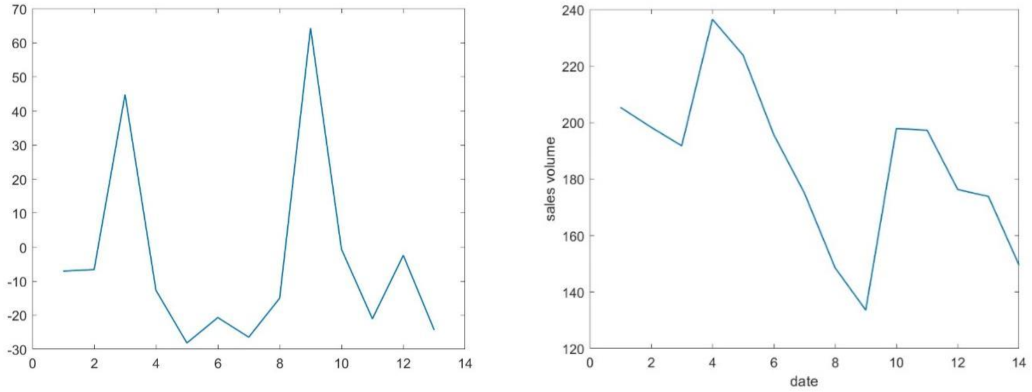

Figure 3: Sales volume raw data and difference reduction plot

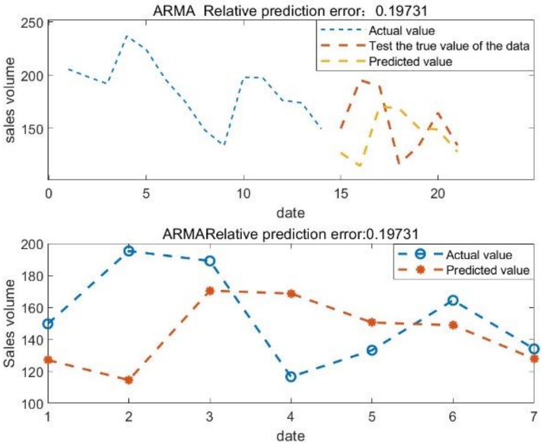

Figure 4: Sales Forecast Chart

After calculating the ACF and PACF, the minimization of information criterion was used to derive a minimum AIC of 6.912 for the foliage category. Predictions were made to obtain the future sales of the foliage category for the next seven days as shown in Table 1 and the relative error of prediction was derived as shown in Figure. 4 as 0.1973. As shown in Figure. 4, the future trend of sales volume will be gradually smooth, in the early period in the lower in the higher fluctuations, but there is no particularly significant fluctuations in general, which indicates that the flower and leafy vegetables the unaffected by other unrelated factors in a certain time under the influence of the sales of stable.

Table 1: Results of sales volume forecast for leafy and flowering vegetables

Date	1	2	3	4	5	6	7
Forecasting	127.03	114.56	170.49	168.69	150.53	148.86	127.84

Subsequently, the original data for the pricing of the foliage category is found to be smooth as shown in Figure 5, and its autocorrelation and partial correlation functions are computed, yielding a minimum AIC of 1.814, and the prediction model yields the pricing of the foliage category for seven days in the future as shown in Table 2, and yields a relative error of the prediction as shown in Figure 6, which is 0.3334

Table 2: Pricing forecast results for leafy and flowering vegetables

Date	1	2	3	4	5	6	7
Forecasting	127.03	114.56	170.49	168.69	150.53	148.86	127.84

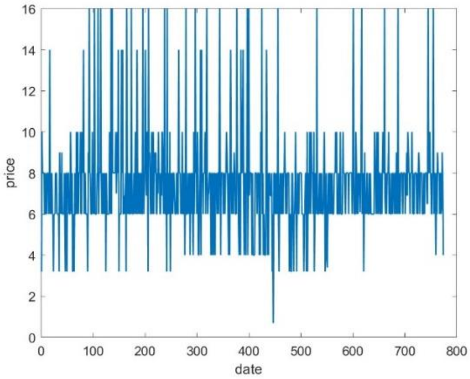

Figure 5: Figures of raw data for foliar species

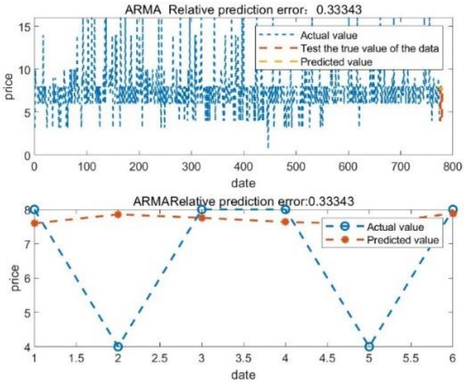

Figure 6: Pricing Forecast Chart

From Figures 5 and 6, it is easy to see that vegetable pricing remains stable without being affected by external circumstances, and although the fluctuations are obvious, they generally show regular ordering. The prediction curve, although not significantly changed, satisfies the error range in short- and medium-term prediction and meets the prediction requirements.

By forecasting the sales volume and cost-plus pricing of leafy vegetables and cauliflower for the next seven days, it is possible to synthesize the sales for the next seven days. It also provides more accurate results for fresh vegetable sales.

4. Conclusions and outlook

This paper summarizes the distribution patterns and interrelationships among vegetable categories using a systematic clustering model, predicts the total daily replenishment of each fresh vegetable category and the vegetable pricing strategy based on cost-plus pricing, and establishes an ARMA model for predicting the total replenishment and pricing of foliage and leafy categories, which yields predicted data for the next seven days.

This paper speculates that if multiple data individual factors such as vegetable item mix data, customer data, replenishment and pricing data of homogeneous category superstores, and dynamic pricing data of vegetable items are added, not only can the market demand be predicted more accurately in a limited space, but also the attrition rate of vegetable items can be minimized so that the profit can be maximized. This further improves pricing and replenishment strategies for fresh vegetables.

In this paper, different algorithms are likely to cluster into chains when performing cluster analysis, making the results of the analysis far from the correct relationship, which requires the need to ensure the accuracy and simplicity of the data when acquiring the data. The ARMA time series forecasting method used in this paper is not effective for pricing and replenishment strategies for different vegetable individual products and categories in long-term forecasting, and the forecast tends to have a large deviation when there are large changes in the outside world. At this time, the establishment of other algorithmic models may achieve better results, such as the Artificial neural network model, Hybrid forecasting method, and so on.

References

[1] Shukla M, Jharkharia S. ARIMA models to forecast demand in fresh supply chains [J]. International Journal of Operational Research, 2011, 11(1): 1-18.

[2] Jha G K, Sinha K. Time-delay neural networks for time series prediction: an application to the monthly wholesale price of oilseeds in India [J]. Neural Computing and Applications, 2014, 24: 563-571.

[3] Dharavath R, Khosla E. Seasonal ARIMA to forecast fruits and vegetable agricultural prices[C]//2019 IEEE International Symposium on Smart Electronic Systems (iSES) (Formerly iNiS). IEEE, 2019: 47-52.

[4] Fan T, Xu C, Tao F. Dynamic pricing and replenishment policy for fresh produce[J]. Computers & Industrial Engineering, 2020, 139: 106127.

[5] Wang L, Feng J, Sui X, et al. Agricultural product price forecasting methods: research advances and trend[J]. British Food Journal, 2020, 122(7): 2121-2138.

[6] Yin H, Jin D, Gu Y H, et al. STL-ATTLSTM: vegetable price forecasting using STL and attention mechanism-based LSTM [J]. Agriculture, 2020, 10(12): 612.

[7] Yoo T W, Oh I S. Time series forecasting of agricultural products' sales volumes based on seasonal long short-term memory [J]. Applied sciences, 2020, 10(22): 8169.

[8] Purohit S K, Panigrahi S, Sethy P K, et al. Time series forecasting of price of agricultural products using hybrid methods [J]. Applied Artificial Intelligence, 2021, 35(15): 1388-1406.