Research On Supermarket Replenishment and Pricing Strategies Based On K-Means Clustering and ARIMA Time Series Model

Qi Liu *, #, Wenbin Cui #, Xinbo Zhang #

Mathematical Sciences, Liaocheng University, Shandong, China, 252000

* Corresponding Author Email: liuqimathlcu@163.com

#These authors contributed equally

Abstract. At present, the shelf life of fresh supermarket vegetables is short, and the quality gradually declines with time. In order to maximize the benefits of supermarkets, this paper established a K-means clustering model, conducted cluster analysis of various categories, explored the correlation between different categories and different items of vegetables, and used ARIMA time series model to predict the pricing and replenishment volume of various commodities in supermarkets, which had important practical significance for supermarket decision-making.

Keywords: Replenishment strategy; pricing analysis; K-means clustering; ARIMA time series model.

1. Introduction

As social productivity increases and market competition intensifies, people are demanding more and more fresh produce in terms of variety and quality. Supermarkets replenish daily based on sales and demand to ensure the freshness of vegetable items, and determine the total amount of replenishment and pricing strategy based on total sales and cost-plus pricing to improve profitability.

Replenishment and pricing strategy has become a research hot spot, the core of which is to predict future supply and demand based on historical and current dynamic data, establish a model to accurately predict future demand, and obtain the optimal pricing method[1]. Ruchir Kulkarni et al, used Sarima model for product sales analysis[2]. Tirta Helmy et al, applied automatic clustering methods and fuzzy logic relationships to manage historical data to obtain forecasts of the number of sales in future periods[4]. Zhang Xinchen et al, applying Wavelet Transform to Smooth the Fluctuation of Price Data and Improve the Accuracy of Price Prediction[5]. Lieke M. van der Heide et al, use a rolling time-domain algorithm for the defined mixed-integer programming formulation to realize dynamically projected needs[6]. Koichi Kurumatani et al, applied ARIMA, RNN, TAPTP, and DFTS models to replenishment and pricing forecasting, respectively, and showed good properties in some aspects[7-11]. This paper innovatively applied K-means cluster analysis and ARIMA time series model to the study of supermarket replenishment and pricing strategy, which provided a new solution to this problem. This scheme provides important market segmentation information for supermarkets, helps supermarket operators to better understand different varieties of vegetables, and formulates targeted commodity replenishment and pricing strategies. It provides guidance for supermarkets to optimize commodity display and promotion strategies, so that supermarkets can attract customers more effectively and increase sales and profits.

2. Mathematical model establishment

2.1. Cluster analysis

According to the historical statistics of the sales unit price and sales volume of each vegetable single product and category of a supermarket, the sample has been divided into K clusters according to the distance between the samples using the unsupervised clustering algorithm of the distance group, and the contour coefficient has been calculated for each point:

$$S(\mathrm{i})=\frac{b(i)-a(i)}{\max(a(i),b(i)}\tag{1}$$

Where a(i) is the average distance of the sample point x(i) to other sample points in the same cluster, which is called the degree of condensation; b(i) is the average distance from sample point x(i) to all samples of the nearest cluster, called the degree of separation.

K-meas clustering generally has four steps:

(1) K points are randomly selected as cluster centers, that is, K class center vectors.

(2) The distance from other sample points to the center vector of each class is calculated separately, and it is divided into the closest class.

This article uses Euclidean distances for calculations:

$$dist(X,Y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}\tag{2}$$

(3) Update the center vector of each class.

(4) If the distance between the newly calculated center vector and the original center vector is less than a certain set threshold, it can be considered that the clustering has reached the expected result and the algorithm terminates.

If the distance between the new center vector and the original center vector changes greatly, several iterations are required until the desired result is achieved.

2.2. ARIMA model establishment

Based on the existing time series data of each vegetable category, the stationarity analysis and the ADF test of the original sequence were carried out. Then, the fitting order difference and white noise test were conducted on the data. After the ADF test of the difference series, the $ARIMA(p,d,q)$ model was established for the original sequence, and the sales volume of the future cycle was predicted by using the existing data.

Model $ARIMA(p,d,q)$ is as follows:

$$\begin{cases}\Phi(B)\nabla^{d}Y_i=\Theta(B)\varepsilon_t,\\ E(\varepsilon_t)=0,Var(\varepsilon_t)=\sigma_\varepsilon^2,E(\varepsilon_t\varepsilon_s)=0,s\neq t,\\ E(\varepsilon_tY_s)=0,s<t.\end{cases}\tag{3}$$

If the above formula satisfies $\nabla^d=(1-B)^d$ in Formula $\varphi_0\neq 0$, the model is the centralized $ARIMA(p,d,q)$ model. Where $\{Y_t\}$ is the time series, $\varphi_0,\varphi_1,\ldots\varphi_p$ is the real number, $\{\varepsilon_i\}$ is the zero-mean white noise sequence, and B is the delay operator.

3. Analysis of experimental results

3.1. Results of cluster analysis experiments

Sample correlation analysis has been carried out using K-means clustering. The optimal number of clusters, k = 2, has been determined using the method of squared intra-cluster deviation and inflection points with contour coefficients. At this time, the contour coefficient is 0.85, and its correlation heat map is shown in Figure 1.

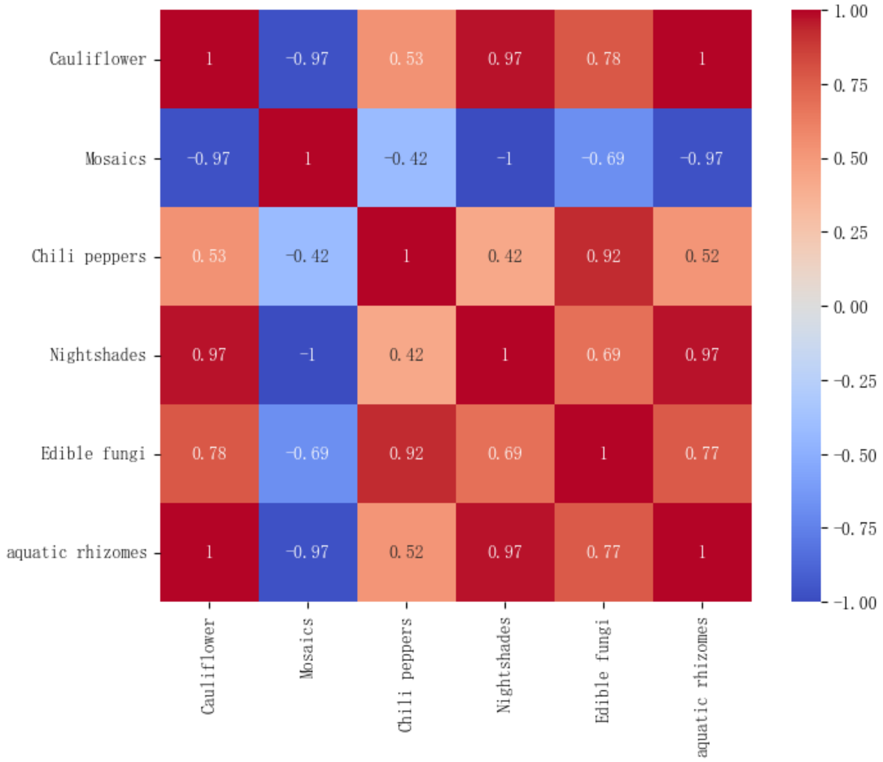

Figure 1. Correlation heat map

As shown in Figure 1, the scale on the right side of the heat map demonstrated how different correlation coefficients corresponded to shades of revealed color. Based on the tightness and degree of dispersion of the contour coefficient with the clustering results, it was found that the correlation between the Brassicaceae class and the aquatic rhizome class was 1, indicating a very strong multicollinearity. The correlation between Brassicaceae and aubergine was higher at 0.97, suggesting the second highest multicollinearity. The lowest correlation between the foliar and aubergine categories was -1, indicating that the two categories were the least correlated. It could be seen that the vegetable category and single product sales were closely linked to people's usual vegetable dietary habits.

3.2. Analysis of ARIMA model prediction results

The calculation shows that the differential sequence is not a white noise sequence. Figure 2 shows the autocorrelation plot after first-order differencing, and Figure 3 shows the partial correlation plot after first-order differencing.

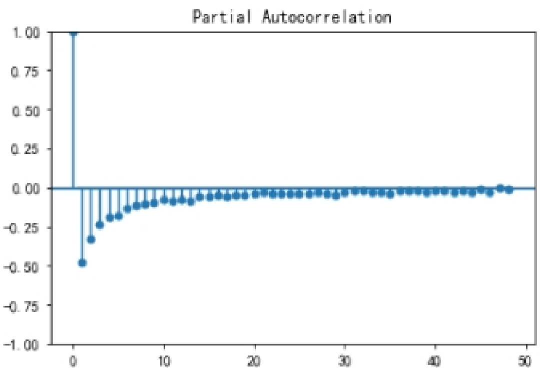

Figure 2. Autocorrelation diagram

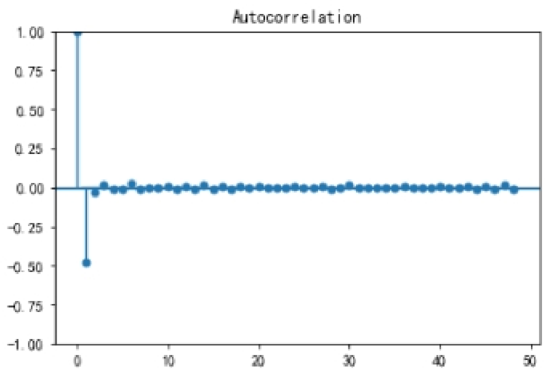

Figure 3. Partial correlation plot

In this paper, the ARIMA model has been ordered, the range of p and q has been selected, and BIC has been checked. When BIC has been taken as the minimum value, p=0, q=0. The model has been used to fit the sequence after first-order difference, and the ARIMA model has been established for the original sequence, and then its parameters have been tested. Finally, we have obtained the forecasted sales volume, i.e. the total amount of replenishment, as shown in Table 1. The pricing strategy has been shown in Table 2.

Table 1 ARIMA model forecasts sales (kg)

	Brassicaceae	Cruciferae	Peppers	Aquatic rhizomatous	Edible Mushroom	Aubergine
July 1	17.2910	139.1524	99.0349	20.8240	59.9435	25.6346
July 2	13.2549	110.7320	89.3218	14.9454	50.5195	30.2553
July 3	12.2644	89.2730	85.1414	12.4953	51.3496	19.5367
July 4	12.9402	101.2819	86.3494	16.0495	48.1561	20.5649
July 5	13.0849	108.7264	60.3245	15.4293	49.3519	15.0456
July 6	14.0974	118.0382	68.4415	17.4255	35.5991	12.4365
July 7	14.0852	130.8302	68.2520	17.8904	40.2564	16.4235

Table 2 ARIMA model prediction pricing (Yuan/kg)

	Brassicaceae	Cruciferae	Peppers	Aquatic rhizomatous	Edible Mushroom	Aubergine
July 1	0.4779	0.6213	0.6359	0.4869	0.5749	0.5732
July 2	0.5102	0.6216	0.5988	0.5131	0.5893	0.5713
July 3	0.5179	0.5766	0.5984	0.5264	0.5896	0.5746
July 4	0.4968	0.6362	0.5968	0.5138	0.5799	0.5698
July 5	0.4985	0.6314	0.5975	0.5017	0.5768	0.5693
July 6	0.4873	0.6257	0.6102	0.4968	0.5773	0.5689
July 7	0.4839	0.6256	0.6034	0.4985	0.5824	0.5687

Table 1 and Table 2 show the sales and pricing of different vegetable categories predicted by the ARIMA model from July 1 to July 7, respectively. As can be seen from Table 1, the sales volume of different vegetable categories fluctuated within a week, and the trend of each vegetable category was different. Pricing strategies for different vegetable categories fluctuated somewhat during the week, but the overall trend was relatively stable. The processing and analysis of historical data by ARIMA model shows that the error between the predicted value and the actual situation is small, but there is still a certain error due to irresistible factors such as weather.

4. Conclusion

Based on K-means clustering and time series model, this paper proposes a comprehensive model, which has advantages in large data set processing, time series prediction and multivariate problems. However, at the same time, it is necessary to note that the multiple linear regression model has high data requirements, the model is sensitive to parameter selection, and is based on the assumption of stationary time series. The comprehensive model provided in this paper provides a more intelligent and comprehensive consideration in the pricing strategy of supermarket vegetable products, which integrates factors such as market demand, competitors, geographical location and supply chain to ensure the freshness of goods and maximize the revenue of supermarkets. In addition, in terms of the promotion of the model, the model has broad potential in the fields of market segmentation, personalized recommendation, anomaly detection and so on. The trained model can be used to predict future data. With the continuous development of artificial intelligence and machine learning technology, the model will be more intelligent and can automate decision-making and resource allocation, so as to achieve a wide range of applications in supply chain management, production planning, traffic management and other fields.

References

[1] Zhou Yue, Shen Xiaobei,Yu Yugang.Inventory control strategy: based on demand forecast error [J]. Modern Supply Chain Research and Applications,2023,5(2):74-101.

[2] Ruchir Kulkarni, Milind Rane.Pattern Recognition - Product Sales Analysis Using SARIMA Model in Time Series Forecasting [J]. International Journal of Science and Research (IJSR), May 2020, 5(9):345-347.

[3] Tirta Helmy,Perdana Novario Jaya,Mulyawan Bagus. Sparepart sales clusterization and prediction using automatic clustering algorithm [J]. IOP Conference Series: Materials Science and Engineering,2020,1007(1).

[4] Jiho Yoon,Srinivas Talluri,Hakan Yildiz. William Ho. Models for supplier selection and risk mitigation: a holistic approach [J]. International Journal of Production Research, 2018, 56(10):3636-3661.

[5] Xinchen Zhang, Linghao Zhang, Qincheng Zhou, Xu Jin. A Novel Bitcoin and Gold Prices Prediction Method Using an LSTM-P Neural Network Model[J].Computational Intelligence and Neuroscience,Volume 2022 ,2022.

[6] Lieke M. van der Heide, Leandro C. Coelho, Iris F.A. Vis, Roel G. van Anholt.Replenishment and denomination mix of automated teller machines with dynamic forecast demands[J].Computers and Operations Research. Volume 114 , 2020. 104828-10482.

[7] Koichi Kurumatani.Time series forecasting of agricultural product prices based on recurrent neural networks and its evaluation method [J]. SN Applied Sciences, 2020,2(8): 801-806.

[8] Foteini Kyriazi, Dimitrios D. Thomakos, John B. Guerard.Adaptive learning forecasting, with applications in forecasting agricultural prices[J].International Journal of Forecasting , 2019, 35(4):1356-1369.

[9] Miaomiao Ji, Peng Liu, Zhao Deng & Qiufeng Wu. Prediction of national agricultural products wholesale price index in China using deep learning [J].Progress in Artificial Intelligence,2022,265(1): 121-129.

[10] Carlos Alberto Orge Pinheiro, Valter de Senna.Multivariate analysis and neural networks application to price forecasting in the Brazilian agricultural market [J]. Ciência Rural,2017,47(1).

[11] Lorenzo-Espejo A, Muñuzuri J, Guadix J, Escudero-Santana A A Hybrid Metaheuristic for the Omnichannel Multiproduct Inventory Replenishment Problem[J]. Journal of Theoretical and Applied Electronic Commerce Research,2022,2(17):476-492.

[12] Kalaiarasan T R, Anandkumar V, Ratheesh Kumar A M.Sales Forecasting using RNN [J], International Journal of Innovative Technology and Exploring Engineering, 2019, 8(9):2748-2751.