A study of fresh produce pricing and replenishment strategy based on FCM cluster analysis and ARIMA-GARCH time series analysis modelling

Miao Wang *

Department of Automation, JiangSu University of Science and Technology, Zhenjiang, China, 212100

* Corresponding Author Email: 15161298895@163.com

Abstract. Aiming at the shortcomings of poor robustness and weak applicability in the formulation of pricing and replenishment strategies of fresh food superstores and the limitations of ARIMA model in pricing strategies, this paper is based on maximizing superstores' revenue, and proposes to establish an FCM cluster analysis model based on the elbow rule to analyse the correlation of different single products, and concludes that the sales volume of edible mushrooms and aquatic roots and tubers have a strong correlation, and spinach and choy sum have a strong correlation. Then the total sales volume of each category was negatively correlated with cost-plus pricing through least squares fitting. Finally, based on the ARIMA-GARCH model, a daily replenishment model was established, and it was predicted that the maximum daily replenishment of foliage and flowers would reach 213.908kg on 2 July, and the minimum daily replenishment of eggplant would reach 8.415kg on 4 July in the coming week.

Keywords: Isolated Forest, Genetic Algorithm, FCM Cluster Analysis, ARIMA-GARCH.

1. Introduction

Fresh agricultural products as one of the important commodities in economic crops, with short freshness period, high preservation requirements, perishable and other characteristics, how to develop and optimize the replenishment and pricing policy has become a key area of concern [1-2], explore timely control and risk avoidance strategy is the current stage of the problem to be solved [3].

Numerous studies have shown that good replenishment and pricing strategies are conducive to fresh produce superstores to have higher cost markups and thus higher profits [5-6]. In addition, external factors such as epidemic risk and extreme weather can reduce product ordering and are detrimental to the profitability of all parties in the supply chain [7-9]. Supermarkets need to replenish products daily based on their historical sales volume and actual market demand, so it is of practical significance to explore the sales volume of different varieties of vegetables at different times to develop replenishment and pricing strategies.

The ARIMA optimisation model can effectively predict and optimise the automatic pricing and replenishment decisions of vegetable goods, which is of great significance in ensuring the normal operation and revenue of fresh produce superstores. However, the ARIMA model itself has the disadvantages of being sensitive to outliers and not applicable to nonlinear processing [10].

In order to study the problem of fresh food replenishment and pricing strategy, open source data is used:http://www.mcm.edu.cn/html_cn/node/c74d72127066f510a5723a94b5323a26.html, which includes product information of six vegetable categories distributed by a supermarket, sales flow breakdown data, wholesale prices of vegetable products, and recent wastage rates of vegetable products. Based on the above data, this paper establishes the FCM cluster analysis model using the elbow rule to analyse the correlation of different single products, and then obtains the negative correlation between the total sales volume of each category and the cost-plus pricing through the least-squares fitting method, and finally establishes the daily replenishment volume prediction model based on the ARIMA-GARCH model to formulate the optimal replenishment plan for the coming week.

2. Correlation analysis modeling

2.1. Assumptions and Justifications

In order to better address the problem, propose the following hypothesis and explain the rationality.

Assumption 1: The warehouse inventory is large enough that there is no shortage of inventory space.

Justification: It is assumed that there is no inventory backlog for ease of calculation.

Assumption 2: Supermarkets are able to fulfil their daily orders without any shortage of supply from their suppliers.

Justification: Suppliers may have problems with untimely supply due to small probability events such as extreme weather, which is ignored here for the purpose of ideal modelling.

Assumption 3: Unit transport costs are the same and pricing does not vary significantly due to transport costs.

Justification: Some fresh products have special requirements for transport conditions [10], and there is a significant difference in transport costs, which is ignored here for the purpose of exploring the general rule.

2.2. Data preprocessing

The data was preprocessed using the isolated forest algorithm to produce the results shown in Figure 1 below.

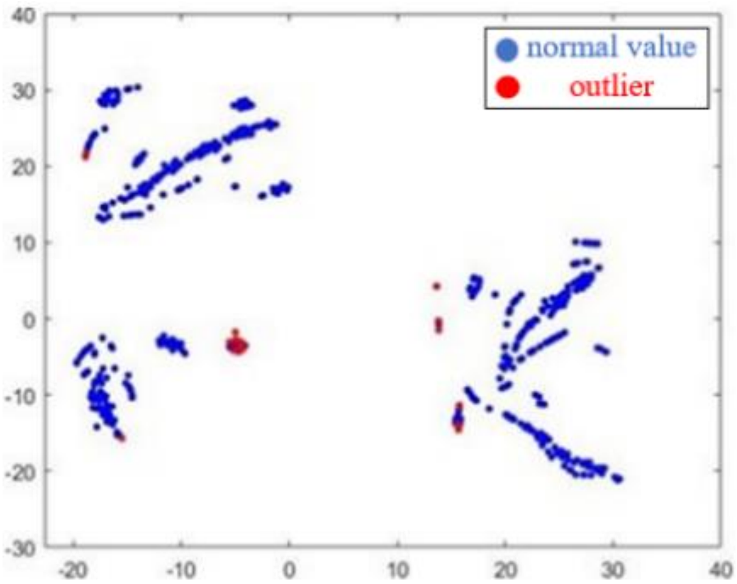


Fig 1. Visualization of the results of the Isolated Forest part of the computation.

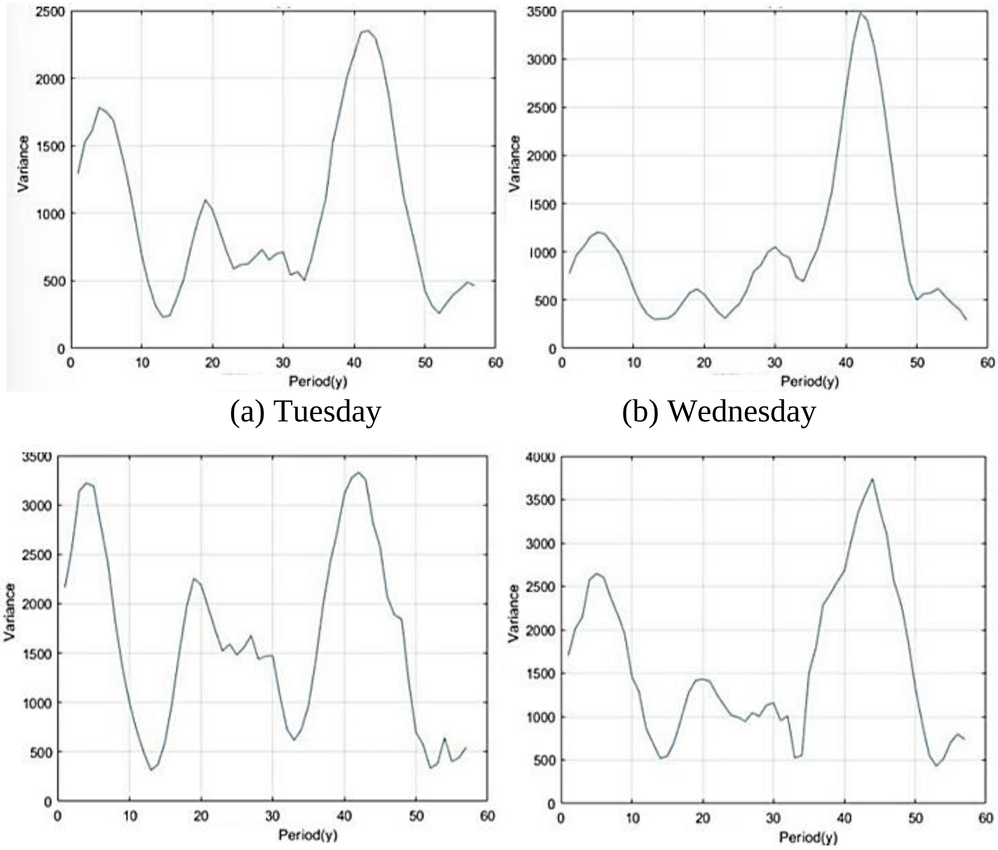
The red points in Fig. 1 are anomalies. 191 groups of anomalous data were excluded from the process of filling in the missing values, and 51 missing data were filled in the process of eliminating the non-missing values.

2.3. Cyclical analysis of sales volume of vegetables in different categories based on wavelet analysis

The wavelet analysis function is shown in equation (1). Where f_s is used as the signal sampling frequency. The original letter f (n) denotes the sum of the components:

$$f(n) = A_1(n) + D_1(n) = A_2(n) + D_1(n) + D_2(n) = A_t(n) + \sum_{i=1}^{t} D_t(n) \tag{1}$$

The Tuesday, Wednesday, Friday and Sunday variance plots of sales volume of flowering and leafy vegetables were obtained as shown in Fig. 2 below.



(a) Tuesday (b) Wednesday

(c) Friday (d) Sunday

Fig 2. ANOVA plot for partial cycle of average monthly sales of foliage and flowers.

From Figure 2 above, it can be seen that the daily variance data fluctuates a lot, which in turn suggests that the sales of leafy and flowering vegetables are somewhat cyclical from Monday to Sunday. By analogy the rest of the categories have a certain periodicity.

2.4. Correlation analysis of sales volume in different categories

By reviewing the relevant literature, there is a correlation between the sales volume of individual items and the time of sale [11-12]. Vegetable commodity categories are more abundant in April-October. In order to establish the Pearson correlation coefficient model for correlation test, first of all, the sales volume of different categories of normal distribution test, for normality test histogram shown in Figure 3 below.

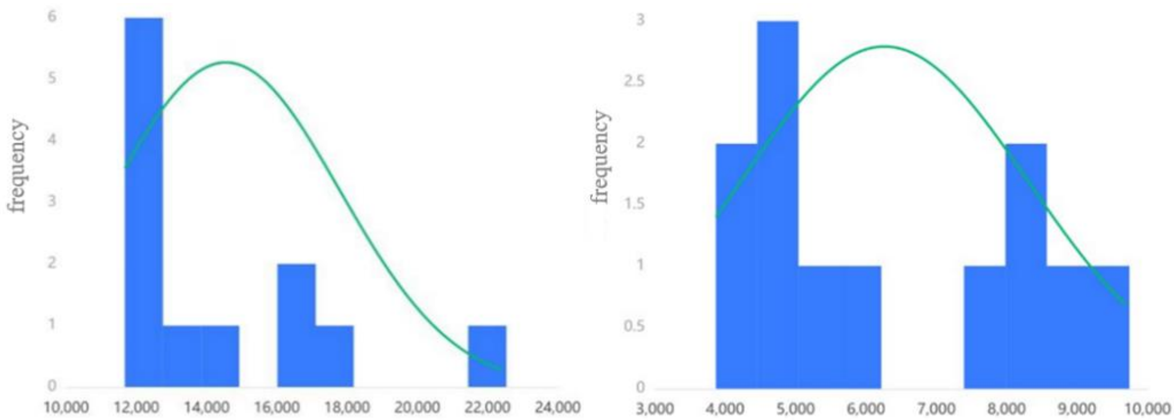


Fig 3. Histogram of normality test (left: sales of foliar species, right: sales of edible mushrooms).

Based on Figure 3, it can be visualized that the sales volume scores of foliage and edible vegetables can be accepted as normally distributed. By analogy, the sales volume of other categories also conforms to a normal distribution.

Record the daily sales volume data of 2 random vegetables as X:{$X_1,X_2,\ldots,X_n$} and Y:{$Y_1,Y_2,\ldots,Y_n$} , with an overall Pearson correlation coefficient of:

$$\gamma_{xy} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \tag{2}$$

Where σ_X is the standard deviation of X:

$$\sigma_X = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})}{n-1}} \quad \sigma_Y = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \overline{Y})}{n-1}} \tag{3}$$

The correlation coefficients of the sales volume of the six vegetables were solved as shown in Table 1 below.

Table 1. Correlation coefficients of sales volume of various vegetables.

	Philodendron	Aquatic rhizomes	Cauliflower	Eggplant	Chilli	Edible Mushroom
Foliage	1	0.298	0.757	0.433	0.376	-0.154
Aquatic Roots	0.298	1	0.543	-0.506	0.465	0.86
Cauliflower	0.757	0.543	1	0.174	0.286	0.266
Eggplant	0.433	-0.506	0.174	1	0.099	-0.708
Chilli	0.376	0.465	0.286	0.099	1	0.349
Edible mushroom	-0.154	0.86	0.266	-0.708	0.349	1

From the above table 1, it can be seen that the sales of edibles and aquatic roots and tubers have a very strong correlation edibles and eggplant sales have a strong negative correlation.

2.5. Analyze the correlation between the sales volumes of different individual products

In order to correlate the sales of each individual product, the elbow rule was first used to determine the number of types needed for the data, which was then verified using the contour method, and then analyzed for specific classifications through hierarchical cluster analysis after being error-free [13].

The core metric of the elbow method is SSE, the sum of squared errors, which is calculated as follows:

$$SSE = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2 \tag{4}$$

Where C_i is the cluster, p is the sample point of C_i, mi is the center of mass of C_i (the value of all the samples in C_i), and SSE is the clustering error of all the samples, which represents the clustering effect. The operation yields the elbow classification map shown in Figure 4 below.

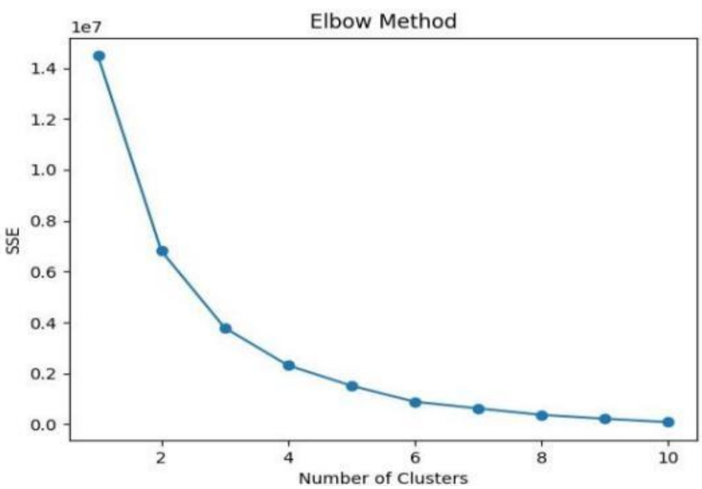


Fig 4. Categorization of the elbow method for all individual items.

According to Fig. 4, it can be seen that according to the horizontal coordinates corresponding to the turning point where the elbow is located can be obtained that all the single items are classified into 4 categories best.

In order to determine the reasonableness of the classification type of all single products, the contour coefficient method is chosen to make a reasonable judgement on the reasonableness [14]. The formula is as follows (5).

$$s(i)=\frac{b(i)-a(i)}{\max\{a(i),b(i)\}},\ s(i)=\begin{cases}1-\frac{a(i)}{b(i)},a(i)<b(i)\\0,a(i)=b(i)\\\frac{a(i)}{b(i)}-1,a(i)>b(i)\end{cases}\tag{5}$$

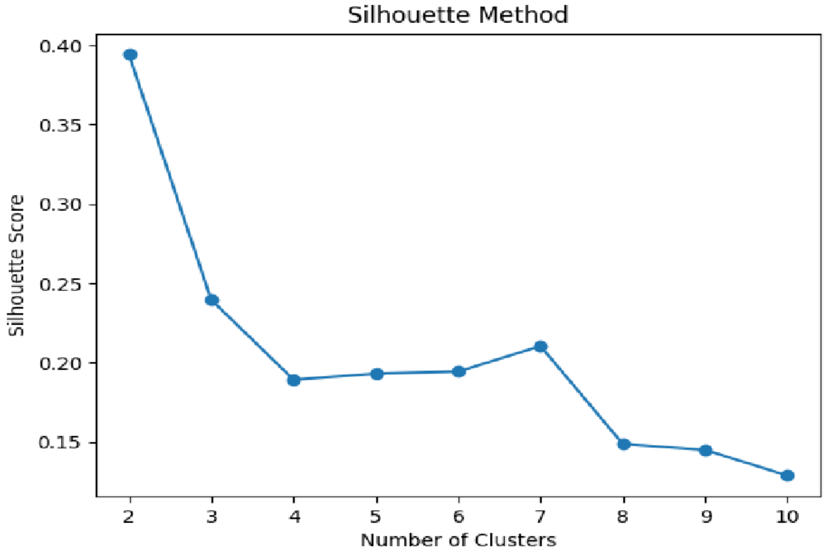


Fig 5. Contour Coefficient Determination Chart.

According to the above formula, the profile coefficient of all single items is 0.3985, as shown in Fig. 5, the classification model is reasonable and good.

2.6. FCM Cluster Visualization

The Euclidean distance between each sample point and the clustering center is calculated to find the clustering center corresponding to the shortest path, and the data objects are merged into the corresponding cluster. The average of the data objects in each cluster is then used as the new clustering center to construct the new class.

$$d(x,v_i)=\sqrt{\sum_{j=1}^{m}(x_j-v_{ij})^2}\tag{6}$$

Where x_j, v_{ij} are the j th attributes of x and a_i.

The FCM clustering was visualized and processed as shown in Table 2 below.

Table 2. Classification of selected individual products.

Category 1	Category 2	Category 3	Category 4
White mushroom(bag)	Golden needle mushroom(bag)	Red bell pepper	Chinese cabbage
Spinach	Screw pepper	Red bell Pepper(portion)	Net lotus root
Chinese cabbage	Green stem scattered flowers	Red bell pepper	Wuhu green pepper
Seafood mushroom(bag)	Shanghai Green	Red bell pepper (portion)	Broccoli
Red Pepper	Xixia shiitake mushroom	Red bell pepper	Yunnan lettuce

3. Prediction model building

3.1. Least squares fitting to analyze the relationship between category sales volume and cost-plus pricing

Let the sample point be (x_i, yi), i=1, 2......., n, and use the least squares method to achieve curve fitting, set the fitted curve as y = kx+b, so that the fitted value.

$$\hat{y}_i = kx_i + b \tag{7}$$

Using the three-year sales volume of each category as the independent variable and cost-plus pricing as the dependent variable, a function was fitted to obtain a linear fit function, as shown in Table 3 below, where *P* is cost-plus pricing and *S* is total sales volume.

Table 3. Fitting functions for each category.

Categories	Function
Foliage	$P = -8.845S + 224$
Cauliflower	$P = -2.029S + 52.21$
Aquatic Roots	$P = -2.895S + 59.68$
Eggplant	$P = -0.02183S + 8.902$
Chilli	$P = -3.167S + 101.6$
Mushrooms	$P = -4.498S + 99.73$

From the above table, it can be seen that the total sales volume of each category is negatively correlated with cost-plus pricing.

Finally, the fit results were tested for superiority. The test results are shown in Table 4 below.

Table 4. Goodness-of-fit test.

Categories	SSE(*10^4)	R^2	Adjustment of the R-square	RMSE
Foliage	139	0.8273	0.8271	35.84
Cauliflower	1.08	0.9807	0.9807	3.157
Aquatic Roots	11.6	0.8911	0.891	10.35
Eggplant	1978	0.9924	0.9924	1.351
Chilli	8.98	0.971	0.971	9.104
Mushrooms	8.4	0.9669	0.9668	8.831

From the above Table 4, it can be seen that the R-square reached 10^4, which is a large value, indicating that the model was fitted better. Among them, eggplant fitted the best, cauliflower and chilli were better.

3.2. Establishment of daily replenishment forecasting model based on ARIMA time series algorithm

The unit root test was performed on the pricing of six categories of vegetables to determine the smoothness of the series. The results of the test for the foliage category are shown in Table 5 below.

Table 5. Unit root test for smoothness.

	Value	Standard error	T	P
Constant	0.510	2.946	0.173	0.862
MA{1}	-0.787	0.038	-20.882	0.000
Variance	10317	345.910	29.825	0.000

From Table 5 above, the MA part of the model satisfies the sufficient condition for reversibility, indicating that the model can be used for prediction and estimation.

Next, the ACF and PACF plots were tested. The ACF and PACF plots for phloem and cauliflower are shown in Figure 6 below.

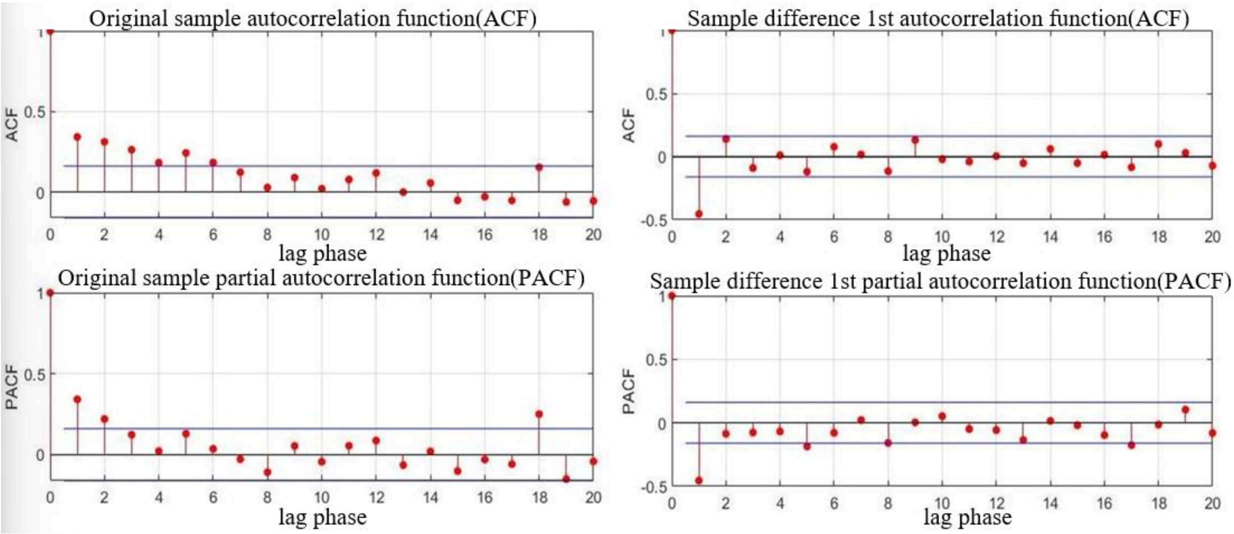


Fig 6. ACF and PACF diagrams for phanerogams and cauliflowers.

As can be seen from Figure 6, the ACF and PACF plots for both cauliflower and foliar categories are trailing, i.e. [15], suitable for ARIMA modelling. By analogy, the same applies to other categories.

Three years of category sales and cost-plus pricing are substituted into the forecasting model, and the established ARIMA-GARCH model is used to forecast the coming week. The resulting daily replenishment totals are shown in Table 6 below.

Table 6. Total daily replenishment of various types of vegetables in the coming week.

Date	Cauliflower	Philodendron	Chilli	Eggplant	Edible mushroom	Aquatic rhizomes
1 July	46.115	192.463	97.632	28.369	50.928	16.054
2 July	25.45	213.908	119.69	43.303	79.626	21.247
3 July	8.083	80.524	72.185	21.224	39.582	10.384
4 July	14.272	127.278	67.12	8.415	38.708	16.454
5 July	16.069	130.182	68.534	15.651	53.742	14.96
6 July	24.367	135.09	89.113	11.511	48.314	22.945
7 July	28.087	130.464	82.286	24.53	39.572	19.419

From the above table 6, it can be seen that the total daily replenishment is maximum in the floral and foliage category and minimum in the aquatic roots and tubers category. It can be seen that the floral and foliage category has the highest daily sales volume and the aquatic rhizome category has the least daily sales volume.

4. Conclusions

In this paper, a Pearson correlation coefficient model based on wavelet transform is constructed to analyse the correlation between different product categories; and the elbow rule is introduced to optimise the FCM cluster analysis to explore the correlation between individual products. The negative relationship between the total sales volume and cost-plus pricing of each category was revealed by least squares fitting. Further, a prediction model for total daily replenishment was constructed using the ARIMA time series algorithm. Based on the above prediction model, combining the ARIMA and GARCH models to predict the replenishment volume, and relying on the objective optimisation model to formulate the pricing strategy, aiming at maximising the revenue of the superstore, which is of significant value in solving the practical problems, the model system is also applicable to the food and beverage outlets' food ingredient ordering scheme and other similar problems. However, the model described in this paper presupposes specific conditions, such as extreme weather conditions, which can disrupt the supply chain and change consumer behaviour, and traffic accidents, which can lead to logistical delays and stable supply from suppliers, which, despite their low probability of occurrence, may still lead to errors in the model, thus restricting the model's general applicability and potential for replication. The next step of the study is to incorporate more potential influencing factors to enhance the robustness and general applicability of the model.

References

[1] Zhu Xiaojing. Research on the development of cold chain logistics of fresh agricultural products under the strategy of rural revitalization [J]. China Storage and Transportation, 2024, (07):52-53.

[2] He Koon. Research on Cost Control of Fresh Food Cold Chain Logistics Based on New Retail Environment of E-commerce [J]. Shanxi Agricultural Economics, 2023, (18):133-135.

[3] Zeng Jiayun,Zhou Taizi. Risk evaluation of fresh food e-commerce supply chain based on community group purchase [J]. China Storage and Transportation, 2024, (05):117.

[4] Li Zheng. Optimization of cold chain logistics of fresh agricultural products [J]. China Logistics and Purchasing, 2023, (15):85-87.

[5] Qin Zhi-dan, Fu Zhi-Yan, LI Jiao. Fresh agricultural supply chain decision considering options and spot market under epidemic risk [J]. Logistics Technology, 2019, 47(13):103-107.

[6] Zhenna Z, Bin Z, Ruijin Y, et al. Recent Developments in the Preservation of Raw Fresh Food by Pulsed Electric Field [J].Food Reviews International, 2022, 38(S1):247-265.

[7] Jinjin H, Min Z, S A M, et al. Technological innovations enhance postharvest fresh food resilience from a supply chain perspective. [J].Critical reviews in food science and nutrition, 2023, 21-23.

[8] Jiang Yingmei, Mou Jinjin. Joint inventory and pricing decision of fresh processed products based on perceived freshness [J]. Journal of Highway and Transportation Science and Technology, 2019, 37(03):151-158.

[9] Zhao J. Pricing and replenishment decision of supermarket vegetables based on optimal profit planning model. National Circulation Economy, 2024, (09):8-12.

[10] Nie Yuxuan. Research on automatic pricing and replenishment strategy of fresh commodities based on ARIMA prediction optimisation model--taking vegetable commodities as an example [J]. Commercial Exhibition Economy, 2024, (05):19-22.

[11] Yang Shuai, HUANG Xiangmeng, WANG Junbin. Research on joint optimization strategy of fresh food shelf distribution and pricing [J]. Supply Chain Management,2022,3(08):49-59

[12] Wang Shaohan, Han Yi, Wang Xiangyu, et al. Two-step ship type classification mapping based on K-Means and XG-Boost algorithms [J]. Journal of Shanghai Marine Transport Research Institute, 2019, 46(03):28-34.

[13] Li Kaian, Zhang Yu, Ren Xuemei, et al. Effect of different grape varieties on aroma of melon-grape complex wine [J/OL]. Journal of Gansu Agricultural University, 2024, 1-16.

[14] Nie Y X. Research on automatic pricing and replenishment strategy of fresh commodities based on ARIMA predictive optimization model -- A case study of vegetable commodities [J]. Trade Fair Economics, 2024, (05):19-22.

[15] ZHAO Yongzhi. Research on the prediction method of tunnel surrounding rock deformation based on LSTM-ARIMA model [J]. National Defence Transportation Engineering and Technology, 2024, 22(04):21-26.