Analysis Of Decision-Making Mechanism for Pricing and Replenishment of Vegetable Commodities Based on Least Squares And ARIMA

Hongyi Li *, Shitong Kang, Yunpeng Shi

Hebei University of Technology, Tianjin, China, 300401

* Corresponding Author Email: 15620588390@163.com

Abstract. This study provides an in-depth study of automatic pricing and replenishment decision-making mechanisms for vegetable commodities based on the relevant data provided. In the research process, firstly, the data were carried out for pre-processing. Secondly, a one-way linear regression model was developed using the least squares method to investigate the relationship between sales volume and cost-plus pricing for each category. The resulting one-dimensional linear regression equations were well fitted, with goodness of fit R2 greater than 0.6. Then, based on the data characteristics of the different categories, appropriate values of q, d and p were selected and a time series (ARIMA) model was built to predict the wholesale price of each category in the coming week. With the final superstore profit as the objective function and the total daily replenishment as the decision variable, an optimization model was constructed for determining the daily replenishment and pricing strategy of each category, and a sensitivity analysis of the model was conducted.

Keywords: Vegetable commodities, Pricing replenishment, least squares, ARIMA prediction, Optimization.

1. Introduction

The development of the economy and the improvement of people's living standards, fresh food superstores have come into people's view, and nowadays it has become a shopping method that people rely on in their lives. The sale of vegetable commodities in the fresh produce superstore has changed the traditional habit of buying vegetables in the farmers' market. The origin of vegetable commodities in the supermarket is clear, and the process of fertilizing and drugging in the production process is under control. After picking accompanied by regulatory processes, there is no secondary spraying to preserve freshness. Vegetable sales in fresh produce superstores thoroughly control the quality of fresh vegetable products. Therefore, the procurement of vegetables in fresh produce superstores has become the main channel for more and more people to buy vegetables.

Vegetable commodities, because of its many varieties, short freshness cycle, price difference between different places of origin, loss rate and other characteristics, resulting in operating profitability difficulties. In order to better serve the public at the same time, increase the profitability of the business, there is an urgent need to study the automatic pricing and replenishment decision-making mechanism of vegetable commodities.

At present, in the vegetable commodity business process, the "cost plus pricing" method to set the sales price. Therefore, the loss in the transportation process and the discount sales caused by the deterioration of commodity quality during the sales process will be included in the cost. Secondly, vegetables are seasonal commodities, and seasonal production and purchases vary greatly. The analysis of market demand and data in the sales process is the key to develop a mathematical model of automatic pricing and replenishment decision-making mechanism for vegetable commodities.

In summary, this study intends to address the following problem: Since vegetable commodities are replenished and purchased on a category-by-category basis, in order to maximize the operational revenue, the total daily replenishment and pricing strategy for July 1-7, 2023, is explored by exploring the potential relationship between the total amount of sales and the cost-plus pricing of different categories of vegetables.

2. Research Methodology

Firstly, preliminary processing of the data was done to get the relevant data of each major category, then the relationship between the sales volume of each category and cost-plus pricing was fitted by MATLAB using the least squares method, after which the data such as the wholesale price of each category in the coming week, i.e., July 1-7, 2023, was predicted by the time-series (ARIMA) model, and finally, the daily replenishment volume was used as the decision variable, and the daily revenue of the superstore was used as the objective function to build an optimization model, and the final result is obtained through Lingo solving.

3. Results and analysis

3.1. Data cleaning

First of all, the data is processed for outliers, and when performing outlier processing, the outlier test method based on normal distribution is discarded because it does not conform to normal distribution. Through observation it can be found that there are obvious logical outliers in the given data. Therefore, outlier processing is mainly focused on logical issues.

Routine testing was not for finding missing values. However, the observation of the total number of days after removing the duplicate values of the days reveals that there are residuals in the total. Therefore, the missing values are found by the method of array comparison.

3.1.1 Logical Problem Outliers

(1) Coded as 102900011034354 data, the number of sales in 2022-06-09 was 160 kg, the product is dumpling leaves (bags) (1), considering the Dragon Boat Festival in 2022 for June 3, the product sales have been over the Dragon Boat Festival of 2022, so it can be identified as anomalous data. The detailed values of the outliers are shown in Table 1:

Table 1. Logical determination of outliers I

Sales Date	Sweeping Sales Hours	Single item coding	Sales volume (kg)	Sales unit price (yuan / kg)
2022-06-09	09:31:57.045	102900011034354	160.000	5.90

(2) A portion of the products have wholesale prices that are far below reasonable wholesale prices [1], such as steak mushrooms with the number 102900005115823 purchased on August 25, 2022, which have a wholesale price of only one penny, which is far lower than the standard wholesale price at a conventional farmer's market. Based on the conventional wholesale prices of related products on the Internet [2], a portion of the products with excessively low wholesale prices were identified as outliers. Detailed data of some outliers are shown in Table 2.

Table 2. logical decision exception value 2

Sales Date	Sweeping Sales Hours	Single item coding	Sales volume (kg)	Sales unit price (yuan / kg)
2022-08-25	102900011021842	0.01	Portabello mushroom	102900005115625
2022-08-26	102900011021842	0.01	Tremella tremella	102900005115748
2022-08-27	102900011007044	0.01	Portabello mushroom	102900005115762
2022-08-27	102900011021842	0.01	Deer mushrooms (box)	102900005115779
2022-08-27	102900051004294	0.01	Tremella tremella	102900005115786
2022-08-28	102900011007044	0.01	Fresh Fungus (1)	102900005115793

3.1.2 Determination of outliers

First, the data were tested for normal distribution to determine whether the data were adapted to normal distribution-based outlier tests such as the 3σ test, and the following is an example of the selling price of lettuce in Yunnan (Table 3).

Table 3. Test of normal distribution of lettuce price in Yunnan

Sample size	Standard deviation	skewness	kurtosis	S-W test	The K-S test
14328	1.207	0.548	0.147	0.901(0.000***)	0.206(0.000***)

Where ***, ** and * represent 1%, 5% and 10% significance levels respectively. The sample size N >= 5000, using the K-S test [3], the significance p-value is 0.000***, the level presents significance and the original hypothesis is rejected, therefore the data does not satisfy normal distribution. The data was further tested for normal distribution using QQ plot [4-7]:

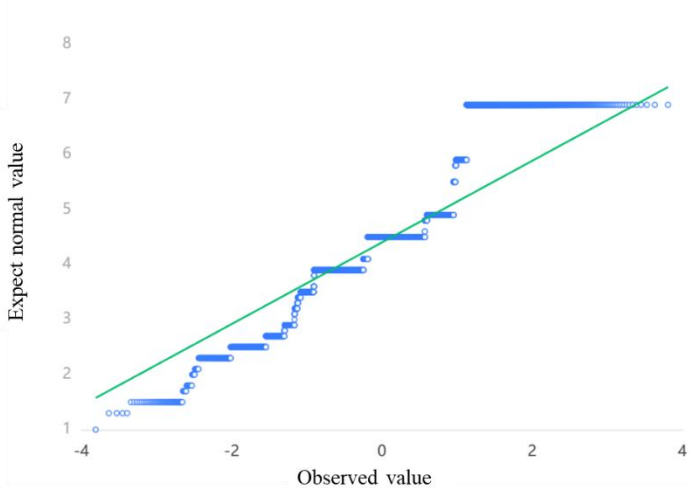

Figure. 1 Wholesale price QMQ chart

The Q-Q plot in Figure 1 shows that the scatter plot of the wholesale price data has a low overlap with the straight line, and therefore it is determined that this data does not conform to a normal distribution. The rest of the data were examined in turn, and it was found that none of them conformed to normal distribution, so the test based on normal distribution was discarded.

The box plot method was discarded because the data were more evenly distributed and some of the individual items had more data at the highest point.

3.1.3 Missing value processing

Observational processing of the collected dataset revealed the presence of missing values. Inspection revealed the presence of ten missing values for sales dates and four for stocking dates. The specific dates are shown in Table 4:

Table 4. date missing value detail value.

Sale date	Purchase date
February 11, 2021	February 11, 2021
February 12, 2021	February 12, 2021
January 31, 2022	January 31, 2022
January 21, 2023	January 21, 2023
November 2, 2022	
November 4, 2022	
November 30, 2022	
December 1, 2022	
December 2, 2022	
December 3, 2022	

Observing the characteristics of the data, the following conclusions are drawn:

(1) February 11, 2021, February 12, 2021, January 31, 2022, and January 21, 2023, are the New Year's Eve and the first day of the New Year in four years, respectively, the store closes with zero sales data, and it is logical that the incoming market closes with zero incoming goods.

(2) November 2, 2022, November 4, 2022, December 1, 2022, December 2, 2022, December 3, 2022, in conjunction with the logic of current events, may be the epidemic caused by the temporary closed management, closed under the closed management of the store closed to stop sales, sales data for zero logical.

Considering the total amount of data is huge, the above data will be deleted.

3.2. Model Establishment and Solution

3.2.1 Data Preprocessing

To analyze the relationship between the total sales volume of each vegetable category and the cost-plus pricing, and to give the replenishment strategy and pricing strategy for the coming week based on the obtained relationship. Considering that the wholesale price and selling price of different individual items within the same category are close to each other, the average of the cost and selling price of individual items within a category on each day is taken as the selling price of this category on that day. The difference in sales volume between different items is too large to use the average value as a substitute, so the total sales volume of each item in a category is used as the sales volume of the category. For the loss rate of the category, the loss rate of different single products varies greatly, and the loss rate of different single products cannot be superimposed on the processing, so in the part of the loss rate, this paper uses clustering to find the typical value of the loss rate of each single product on behalf of the loss rate of the category.

3.2.2 One-dimensional linear regression modeling

After reviewing related information, for sales volume and price, it can be approximated that the two become a linear relationship, so between the sales volume and price of each category, it can be fitted based on the least squares method using a one-way linear regression equation [8]. In this paper, the polyfit command in MATLAB is used for fitting, and the fitting results for each category are shown below. From Figure 2 and Table 5, it can be seen that linear regression fitting yields equations with high goodness of fit and credible results.

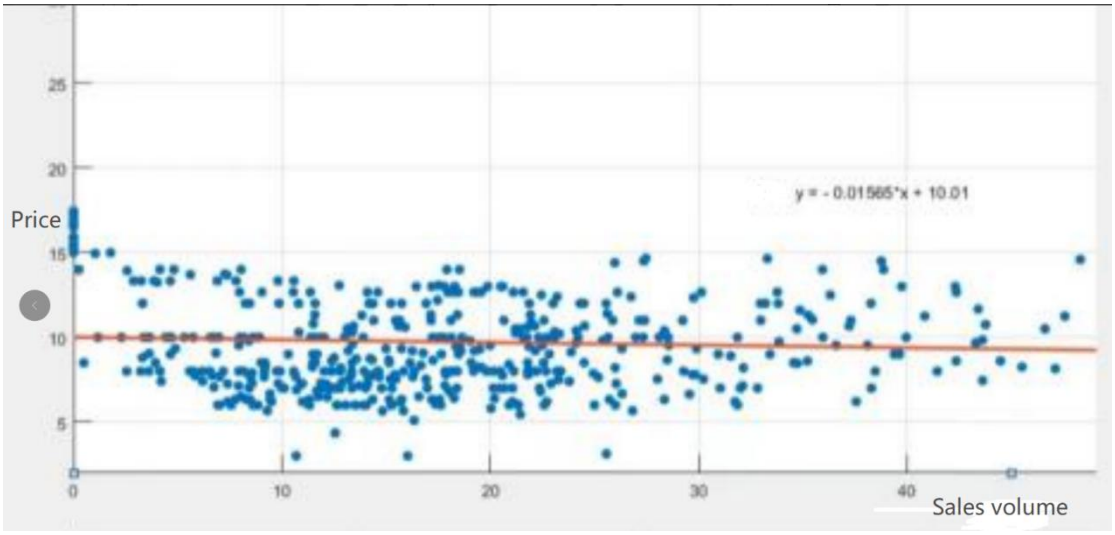

Figure. 2 (a)

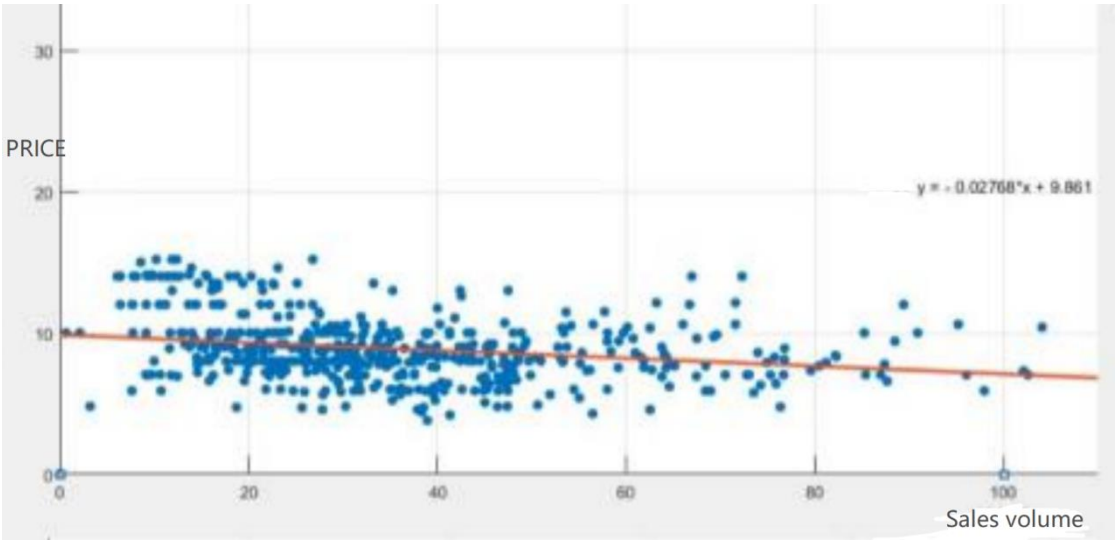

Figure. 2 (b)

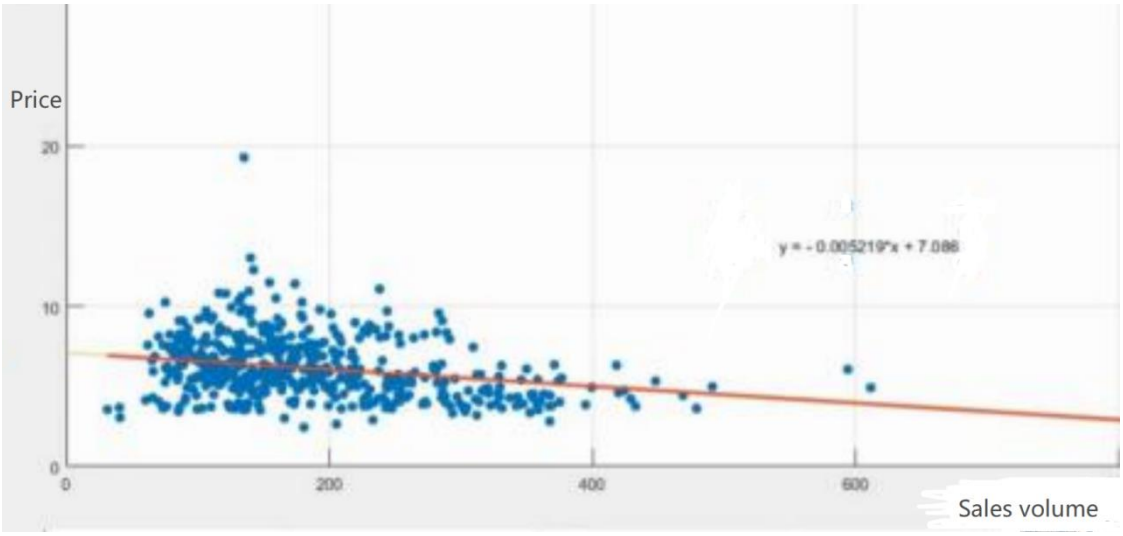

Figure. 2 (c)

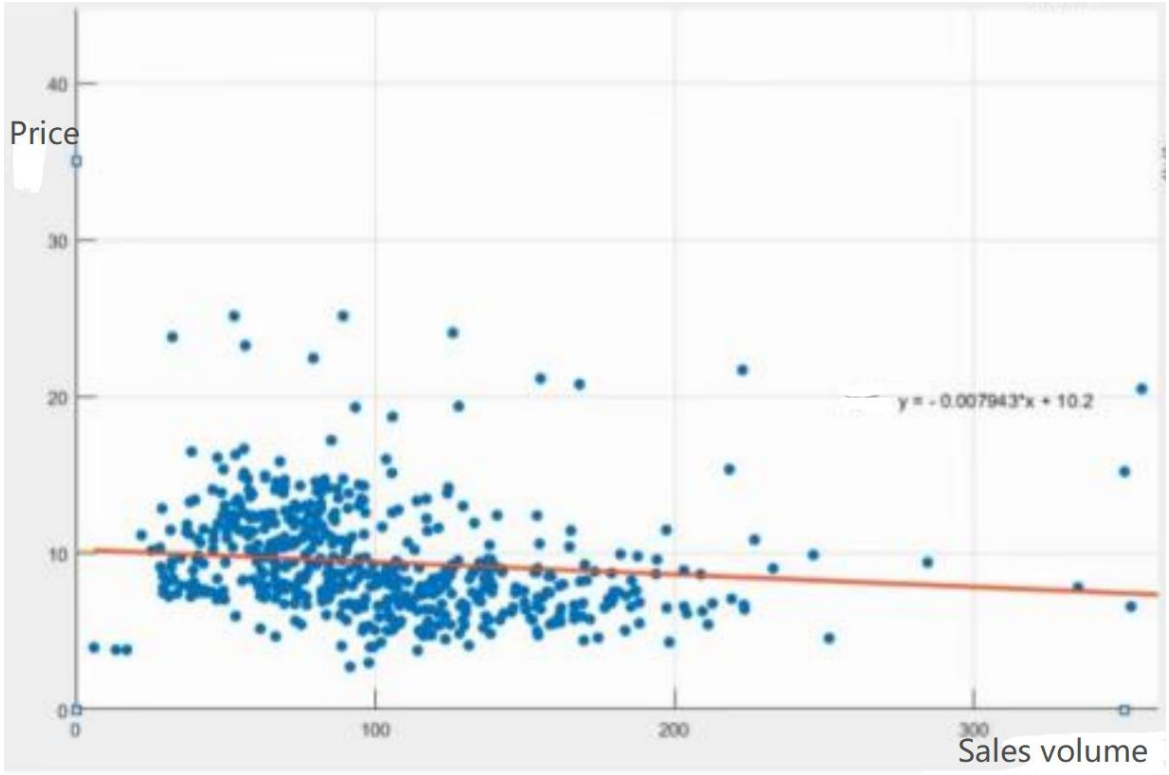

Figure. 2 (d)

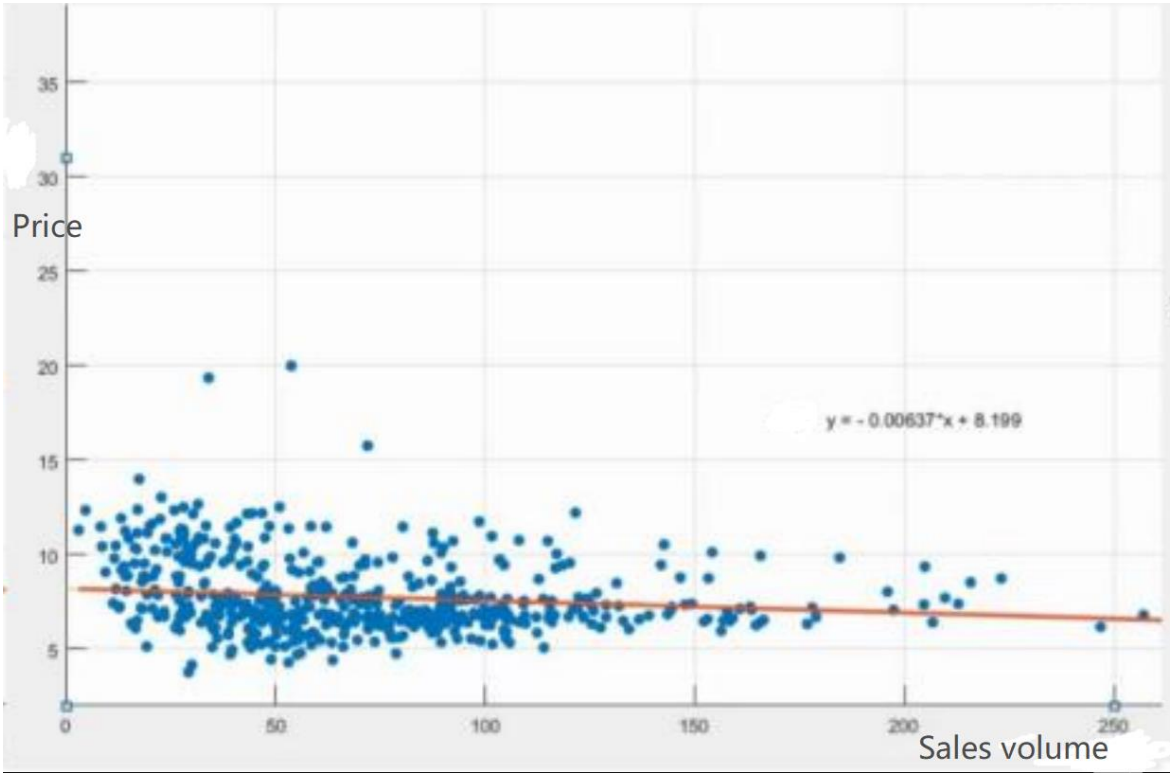

Figure. 2 (e)

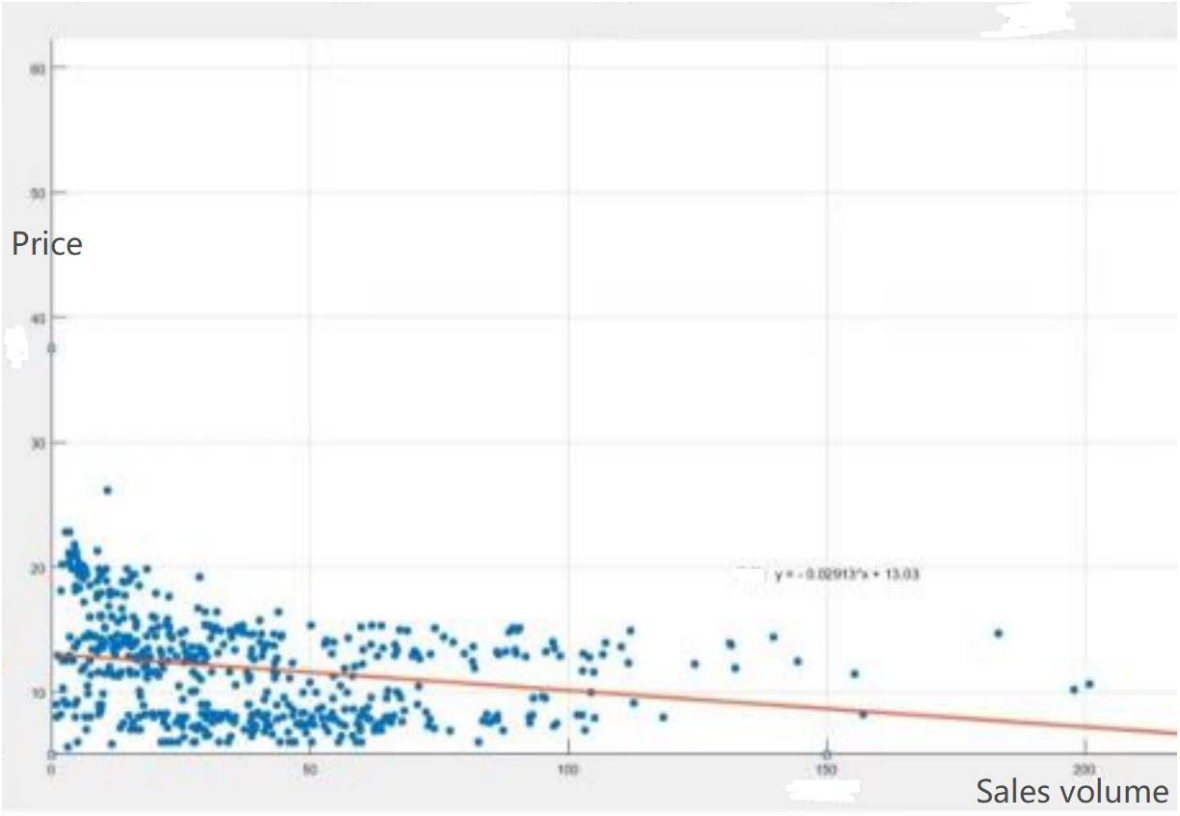

Figure. 2 (f)

Figure. 2 fitting result of univariate linear regression between sales volume and price of each category.

Table 5. fitting results of each category

category	Fitting equation	Goodness of fit R^2
cauliflower	y=-0.02768*x+9.861	0.71
anthophyllum	y=-0.005219*x+7.086	0.66
capsicum	y=-0.007943*x+10.2	0.62
solanacea	y=-0.01565*x+10.01	0.65
Edible mushroom	y=-0.00637*x+8.199	0.81
Aquatic rhizome	y=-0.02913*x+13.03	0.63

3.2.3 Time series analysis (ARIMA)

In order to formulate the replenishment strategy and pricing strategy for the coming week, i.e., to maximize the profit of the superstore by planning the replenishment volume of each category, the wholesale price of each category for the coming week needs to be obtained. From the objective facts, it is known that the wholesale price of vegetables is greatly affected by time, so it can be considered that the wholesale price is a time series, which can be analyzed using the ARIMA model [9,10].

(1) Model principle

ARIMA (Autoregressive Moving Average Model with Difference) is a time series analysis method for modeling and forecasting time series data. The ARIMA model is based on the following three main components: autoregressive (AR), difference (I) and moving average (MA). The basic idea is that over time, the data series formed from forecasts is considered a time series and the model can be used to approximate the series.

The autoregressive (AR) part, denoted as AR(p), where p is the autoregressive order. The autoregressive component takes into account the effect of values at past time points in the time series on current values. Specifically, the AR(p) model uses the values at the past p time points (lagged terms) to predict the values at the current time point. This means that the current value is a linear

combination of past values, where the weight of each past value is determined by the model parameters.

Difference (I) component: denoted as I(d), where d is the difference order. Differencing is used to deal with non-stationarity in the time series. Non-stationarity implies that the statistical characteristics of the time series (e.g. mean and variance) change over time. With the difference operation, a non-stationary time series can be transformed into a stationary time series, making it easier to model. The difference operation calculates the difference between the current value and the previous point in time, i.e., a first-order difference, which can then be differenced several times in succession to achieve smoothness.

Moving Average (MA) component: denoted as MA(q), where q is the moving average order. The moving average component takes into account the effect of the random error term at past time points on the current value. Similar to autoregression, the MA(q) model uses a linear combination of error terms (lagged errors) from the past q time points to predict the value at the current time point. The weights of these error terms are determined by the model parameters.

The general idea of the ARIMA model is to transform the time series data into a smooth series and then use autoregressive and moving average models to model the dynamic structure of the data. The order of the model (p, d, q) is usually chosen by analyzing the autocorrelation function (ACF) and partial autocorrelation function (PACF) of the time series data.

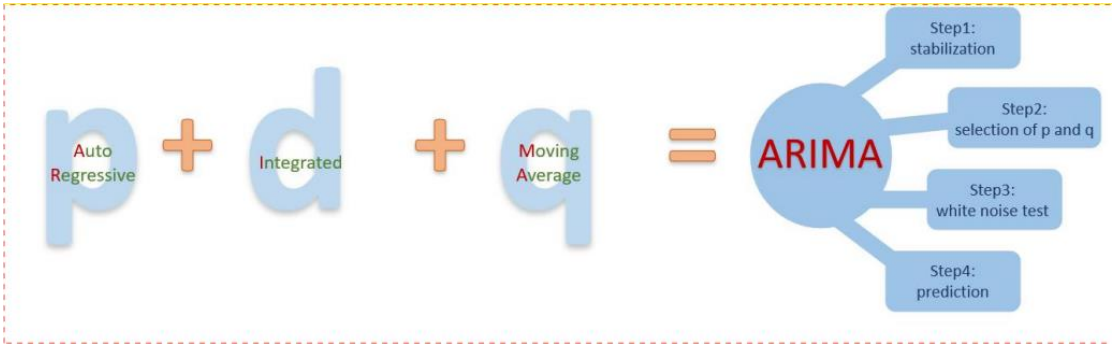

Figure. 3 steps of ARIMA modeling

The general idea of the ARIMA model is to transform the time series data into a smooth series and then use autoregressive and moving average models to model the dynamic structure of the data (Figure 3). The order of the model (p, d, q) is usually chosen by analyzing the autocorrelation function (ACF) and partial autocorrelation function (PACF) of the time series data. The ARIMA prediction model can be written as follows:

$$\hat{P}^{\{t\}} = P_0 + \sum_{j=1}^{p} \gamma_j p^{\{t-j\}} + \sum_{j=1}^{q} \theta_j \varepsilon^{\{t-j\}} \quad (1)$$

Here P is the order of the autoregressive model (AR), q is the order of the moving average model (AM), ε{t} is the error term between time ranging between t and t- 1, γj and θj are the fitting coefficients, P0 is the constant term.

(2) ARIMA (p, d, q) parameter selection

Take the cauliflower class as an example: firstly, we observe whether the data are smooth or not by plotting the sequence diagram. As can be seen from the sequence diagram 4, the smoothness of the original time series is poor, while the smoothness is better after the first-order difference, so the parameter d (difference order) in the ARIMA model is 1. Then, the autocorrelation diagram (ACF) and the partial correlation diagram (PACF) are plotted to determine the specific values of the parameter p (autoregressive order) and the parameter q (moving average order) (Fig. 5). After analyzing the data, it can be seen that p=1, q=1. Similarly, other types of data were analyzed, and the results obtained for each parameter are shown in Table 6:

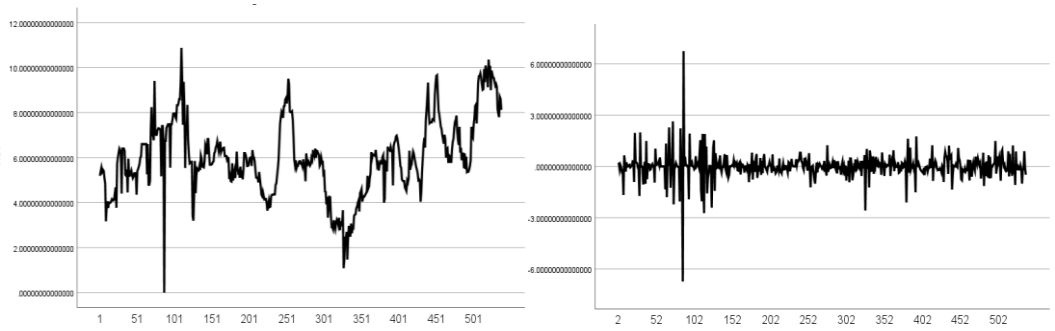

Figure. 4 sequence diagram of wholesale price of cauliflower (before and after difference)

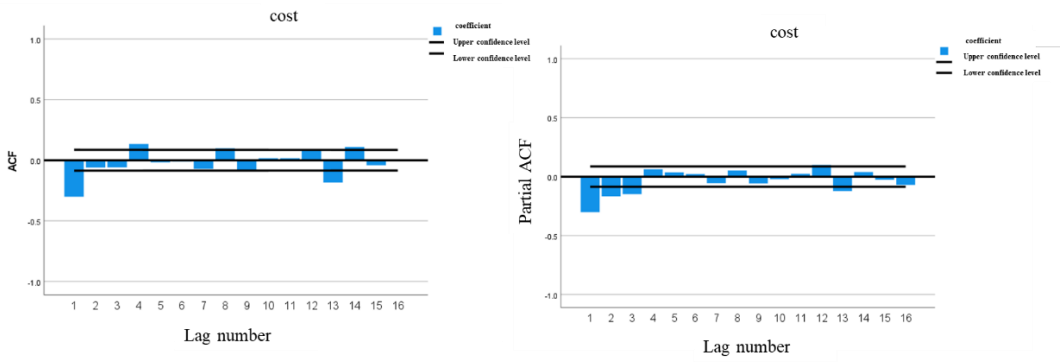

Figure. 5 ACF / PACF diagram.

Table 6. ARIMA model optimal fitting p d q parameter table

	cauliflower	anthophyllum	capsicum	solanacea	Edible mushroom	Aquatic rhizome
p	1	3	1	0	3	0
d	1	1	1	1	0	1
q	1	1	1	2	3	1

(3) ARIMA (1,1,1) modeling

The ARIMA (1,1,1) model is solved by MATLAB and SPSS, and the wholesale price in the coming week is predicted, and the results are shown in Figure. 6 and Table 7 below.It can be seen from the results that the model fits well, and it can predict the wholesale price in the coming week more accurately, and the prediction results are shown in Table 8 below.

Figure 6. Fitting results of cauliflower ARIMA model

Table 7. ARIMA fitting results.

Fitting statistics	Stationary R2	R2	RMSE	MAPE	MaxAPE	MAE	MaxAE	Normalize the BIC
Mean value	0.121	0.832	0.691	7.659	215.057	0.425	6.917	-0.715

Table 8. Forecast of wholesale price of cauliflower in the coming week

model	2023-07-01	2023-07-02	2023-07-03	2023-07-04	2023-07-05	2023-07-06	2023-07-07
Cost forecast	8.2915	8.2971	8.3027	8.3083	8.3139	8.3195	8.3251

3.2.4 Replenishment Volume and Pricing Strategy Solution

From the results obtained in 3.2.3, it can be seen that the sales volume of each category is negatively correlated with the selling price. And because this paper assumes that the volume of the replenished goods is sold in full except for some losses, i.e.:

$$X' = \frac{X}{1-\delta} \tag{2}$$

Where: X': replenishment; X: sales; δ: wastage rate.

From this we can see:

$$W = X'*(1-\delta)*Y - X'*\mu \tag{3}$$

Where: w: profit; μ: wholesale price.

This optimization model is solved using Lingo, and the results are shown in Figure. 7 below for the replenishment quantity of 2023-07-01, for example:

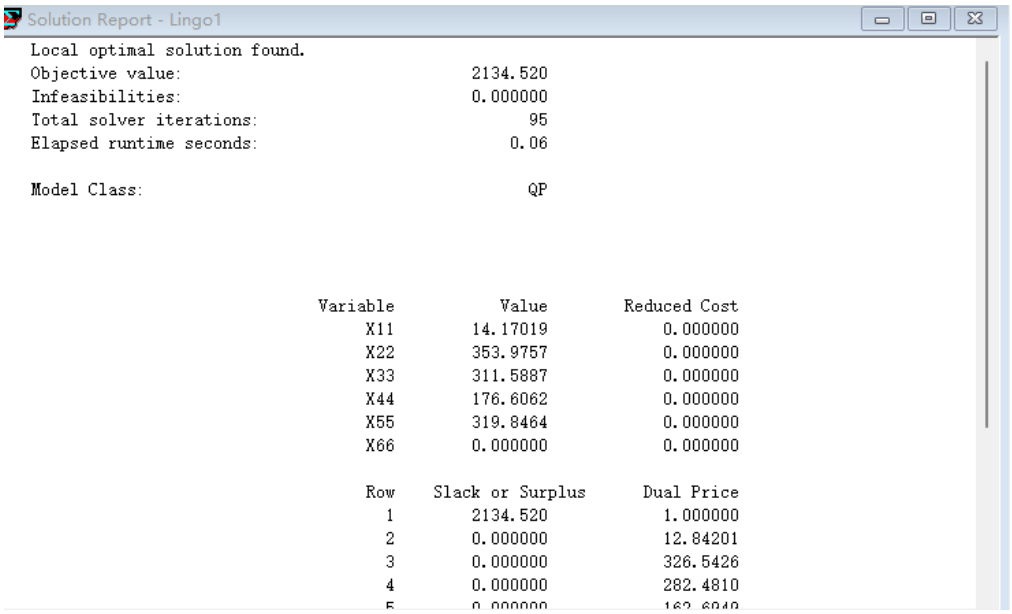

Figure 7. 2023-07-01 replenishment results of each item

To get the maximum revenue, the replenishment volume for the next seven days with the pricing strategy for each category is shown in Table 9 below:

Table 9. Replenishment volume and pricing strategy for each category in the coming week

	cauliflower		anthophyllum		capsicum	
	Replenishment volume	Price	Replenishment volume	Price	Replenishment volume	Price
2023-07-01	14.17019	9.505533	353.9757	5.381774	311.5887	7.956254
2023-07-02	11.86608	9.563333	348.4702	5.408281	290.7386	8.106395
2023-07-03	13.92578	9.511664	349.3071	5.404252	308.6405	7.977483
2023-07-04	13.81906	9.514341	332.1239	5.48698	291.7521	8.099096
2023-07-05	14.3517	9.50098	330.2017	5.496235	253.3566	8.375582
2023-07-06	12.68317	9.542836	337.7051	5.460109	276.9643	8.205583
2023-07-07	12.79775	9.539961	338.0736	5.458335	293.2998	8.087952
	solanacea		Edible mushroom		Aquatic rhizome	
	Replenishment volume	Price	Replenishment volume	Price	Replenishment volume	Price
2023-07-01	176.6062	7.463824	319.8464	6.370414	0	13.03
2023-07-02	169.2348	7.5701	291.1079	6.534714	0	13.03
2023-07-03	161.7813	7.677558	268.2381	6.665463	0	13.03
2023-07-04	176.9508	7.458856	288.8732	6.54749	0	13.03
2023-07-05	184.5735	7.348958	307.5812	6.440535	0	13.03
2023-07-06	176.3104	7.46809	281.566	6.589266	0	13.03
2023-07-07	191.4299	7.250108	268.1664	6.665873	0	13.03

3.2.5 Sensitivity test

In order to assess the stability and reliability of the optimization model, the model parameters in this paper were subjected to sensitivity analysis. From the previous results, it can be seen that the cauliflower category has the highest goodness of fit, and the eggplant category has the lowest goodness of fit, so the cauliflower category and eggplant category do the sensitivity test respectively. Taking cauliflower category as an example, firstly, the cost of cauliflower obtained from the prediction was used to find the replenishment quantity of cauliflower category, and then the cost wholesale price was changed to calculate the replenishment quantity of cauliflower category. By comparing the difference between the original cost wholesale price and the modified cost wholesale price, we can understand the sensitivity of the model to changes in the input data. The results of the analysis are shown in Tables 10 and 11 below, respectively:

Table 10. Optimal purchase quantity of cauliflower disturbance cost

	-10%	-5%	Initial	5%	10%
2023-07-01	17.1	16.326	14.170	12.422	11.256
2023-07-02	14.902	13.134	11.866	11.066	10.973
2023-07-03	16.771	15.007	13.925	12.986	11.582
2023-07-04	15.705	14.836	13.819	13.043	12.382
2023-07-05	16.079	15.503	14.351	13.051	11.985
2023-07-06	14.098	13.045	12.683	11.579	10.981
2023-07-07	15.067	13.674	12.797	11.998	11.065

Table 11. Optimal purchase quantity of eggplant disturbance cost

	-10%	-5%	Initial	5%	10%
2023-07-01	181.023	177.245	176.606	175.348	173.231
2023-07-02	173.246	171.334	169.234	167.253	165.917
2023-07-03	163.357	162.079	161.781	160.786	159.961
2023-07-04	179.837	177.023	176.950	175.312	173.281
2023-07-05	187.231	185.099	184.573	183.657	181.997
2023-07-06	179.269	177.425	176.310	173.241	171.980
2023-07-07	194.356	191.978	191.429	190.321	188.542

4. Conclusions

The present research conducts an extensive examination of the automated pricing and restocking strategies for vegetable products utilizing the provided data. The methodology involved began with preprocessing the data. Subsequently, a univariate linear regression model, derived via the least squares technique, was formulated to examine the correlation between sales volume and cost-plus pricing for individual categories. The resulting linear regression equations were notably precise, with R2 values surpassing 0.6. Following this, considering the varying data attributes across categories, optimal values for q, d, and p were chosen, and an ARIMA time series model was established to forecast the wholesale prices for each category in the subsequent week. Lastly, an optimization model was developed with the ultimate supermarket profit as the objective and the aggregate daily restocking as the variable of choice, aiming to ascertain the daily restocking and pricing tactics for each category, in addition to performing a sensitivity analysis on the model.

References

[1] Cui Xiuhong, Zhang Jiangwei. Vegetable prices have risen, grain and oil prices remain stable: Comments on Beijing Wholesale Price Index of Agricultural Products in December 2020 [J]. Price Theory and Practice,2020(12):160.

[2] https://www.cnhnb.com/hangqing/

[3] Kourosh S M, A. S F. Selecting representative geological realizations to model subsurface [formula omitted] storage under uncertainty [J]. International Journal of Greenhouse Gas Control,2023,127.

[4] Yuan K,Liu H,Han Y. Differential Item Functioning Analysis Without A Priori Information on Anchor Items: QQ Plots and Graphical Test [J]. Psychometrika,2021,86(2).

[5] Wang Ya Nan, Liu Chang. Spearman: From the establishment of the two-factor Theory of Intelligence to the breakthrough in methodology [J]. Journal of Nanjing Normal University (Social Sciences Edition),2011(06):95-101.

[6] He Ling, Wu Lingda, CAI Yichao. Application Research of Computers,2007(01):10-13. (in Chinese)

[7] http://www.nmc.cn/publish/observations/hourly-temperature.html

[8] Fatemeh A, Sanaz S, Taha K, et al.Roadmap for outlier detection in univariate linear calibration in analytical chemistry: Tutorial review[J].Journal of Chemometrics,2022,37(1).

[9] Lei Yu, Zhao Danning, CAI Hongbing et al. Singular spectrum analysis and integrated ARIMA model to predict day long change [J/OL]. Journal of wuhan university (information science edition): 1-17 [2023-09-10].

[10] Xiaohui Y, Zezhong H, Riying X, et al.Optimisation and analysis of an integrated energy system with hydrogen supply using solar spectral beam splitting pre-processing[J].Energy,2024,287.