Research on Replenishment Strategy of Vegetable Commodities Based on ARIMA, Random Forest and Optimization Algorithm

Jiayi Liu[1,*], Yi Zhong[2], Tao Yang[2]

[1]School of finance and economics, Wuhan College, Wuhan, China, 430212

[2]School of information engineering, Wuhan College, Wuhan, China, 430212

*Corresponding author: 2695929852@qq.com

Abstract. With the improvement of social productivity and the intensification of market competition, urban residents pay more and more attention to fresh vegetables and other fresh products. Therefore, this paper investigates the supply demand, pricing and replenishment strategies of vegetable commodities. First, this paper uses ARIMA model to predict the wholesale price of vegetable commodities. Then, this paper uses the random forest model to estimate the sales volume of vegetable commodities. Both models showed great accuracy on the test set. Finally, based on the model results, an optimization model for developing replenishment strategies was constructed and solved using relevant market information. This study helps to help the fresh produce wholesale industry develop more effective sales strategies in advance and reduce resource waste.

Keywords: Optimization model; ARIMA; time series model; random forest; vegetable commodity pricing.

1. Introduction

With the improvement of social productivity and the intensification of market competition, the type and quality requirements for fresh vegetables and other fresh products are increasing. Vegetable price fluctuations affect both producers and consumers, and have a bearing on social livelihood issues.[1] In recent years, the prices of some vegetable markets have shown a trend of increased volatility, and the speed of upward and downward transitions has accelerated significantly. [2] Since the 21st century, in Guangdong Province, for example, with the accelerated growth of income of urban and rural residents in Guangdong, the per capita consumption structure of agricultural products has accelerated to change, and the total consumption of agricultural products has continued to expand, and the total consumption of agricultural products in urban and rural areas of Guangdong ranked first in the country. In 2013, the residents of Guangdong consumed a total of 15,828,000t of food, 10,378,000t of vegetables, and 3,481,000t of meat, poultry, and their products. [3] However, vegetable commodities suffer from the problems of being perishable and difficult to store and may lead to a decline in demand as prices rise. Therefore, price setting and replenishment strategies for commodities become crucial. The ARIMA model combines the dependence of economic phenomena on time series and stochastic fluctuations in economic forecasting, and therefore possesses significant accuracy in capturing short-term economic trends, making it one of the most widely adopted methods in recent years.Based on historical data, Srikanth Sankaran has forecasted the daily demand of fresh vegetable products (onion) in Mumbai wholesale market. Among the models developed and tested, the Seasonal Autoregressive Integrated Moving Average (SARIMA) model outperforms other competitors in terms of forecasting accuracy for in-sample and two out-of-sample data sets. The results showed that the model can be used to forecast with a mean absolute percentage error (MAPE) of 14 %, which is considered acceptable for random demand products such as fresh vegetables.[4] Peng Hongxing et al. established ARIMA prediction model, BP neural network prediction model and LSTM neural network prediction model based on time series, analyzed and predicted vegetable prices by three models, and finally compared the experimental results of the three prediction models. The results show that ARIMA model is more accurate than LSTM and BP model in predicting vegetable

prices. [5] Cao Xinyue et al. used X12 - ARIMA and LSTM combination model to forecast urban vegetable prices. [6] It is easy to see from the above studies that it may be difficult to deal with non-linear, non-stationary or irregular time series data by using ARIMA model, especially when the demand is affected by external factors (e.g. weather, holidays, etc.). Moreover, the above studies are susceptible to poor data quality and the presence of missing values and outliers in the data, which have an impact on ensuring the accuracy of the model. For example, Teng Yongping used the X12 seasonal adjustment method and the ARIMA model to conduct an in-depth study on the changes in China's corn prices in the coming period.[7]The results of the study show that ARIMA model has good performance in short-term performance, but not suitable for medium and long term forecasting.[8] Therefore, this paper proposes a random forest model that establishes the relationship between various other aspects such as price and sales volume, and then uses the ARIMA model to predict the price trend of each vegetable category, and finally builds an optimization model to determine the pricing and replenishment strategy when maximizing profits.

2. Methods

2.1. Introduction to Modeling Methods

ARIMA Model (Autoregressive Integrated Moving Average): this is a classical model used for time series analysis, forecasting and control.[9] Autoregressive Integral Moving Average (ARIMA) is a smooth time series forecasting model based on linear regression.[10]The ARIMA model combines the concepts of Autoregression (AR) and Sliding Average (MA) with Integration operation to deal with non-stationary time series. The labeled values at a point in time are affected both by the labeled values over the past period and by chance events over the past period. The AR part is used to handle the autoregressive part of the time series which considers the effect of observations over the past number of periods on the current values and the MA part is used to handle the moving average part of the time series which considers the effect of past forecast errors on the current values.

Random Forest Model: Random Forest is a machine learning algorithm based on integrated learning, where the underlying structure is a number of decision trees that perform classification and regression tasks by constructing multiple decision trees and combining their predictions.[11] Random Forest algorithm can be used for both classification problems and regression problems.[12] The key idea of Random Forest is to enhance the generalization ability of a single decision tree by introducing randomness and to reduce the variance of the prediction by integrating the prediction results of multiple decision trees. Random Forest is a model that combines multiple regression trees and is widely used in classification studies. Random forest regression is able to effectively analyze nonlinear, data with covariance and interactions better than multivariate linear regression and does not require pre-given assumptions about the form of the model, and the regression results are better than regression trees.

2.2. Relationship between total sales and cost-plus pricing for each vegetable category

First, this paper utilizes Equation 1 to derive the cost margin for each individual item at different times.

$$r_i^n = (q_i^n - p_i^n) / p_i^n \tag{1}$$

r_i^n denotes the cost margin of item I of Category N, q_i^n denotes the selling price of Item I of Category N, and p_i^n denotes the wholesale price of Item I of Category N. In this article, we first need to perform an aggregation operation on the sales volume of an item, and then perform an aggregation operation to add up the sales volume of individual items, as a single item may have multiple transaction records at different points in time. After aggregation, the total sales volume data for individual items at each point in time can be obtained. With the total sales figures, it is possible

to calculate the total profit of the following companies w_i^n at different points in time. The specific formula used is shown in Equation 2.

$$w_i^n = v_i^n * (q_i^n - p_i^n) \tag{2}$$

However, as time changes r_i^n, q_i^n, p_i^n and w_i^n also change. Therefore, faced with an unknown data distribution, this paper uses the Spearman correlation coefficient to analyze the correlation between the cost margin of each individual item and other sales. However, in category correlation analysis, it is necessary to aggregate the individual items in different categories rather than simply using the mean-merge method. For total sales and total profit, summation should be performed. The exact formula is shown in Equation 3.

$$\begin{cases} v^n = \sum_{i \in n} v_i^n \\ w^n = \sum_{i \in n} w_i^n \end{cases} .... \tag{3}$$

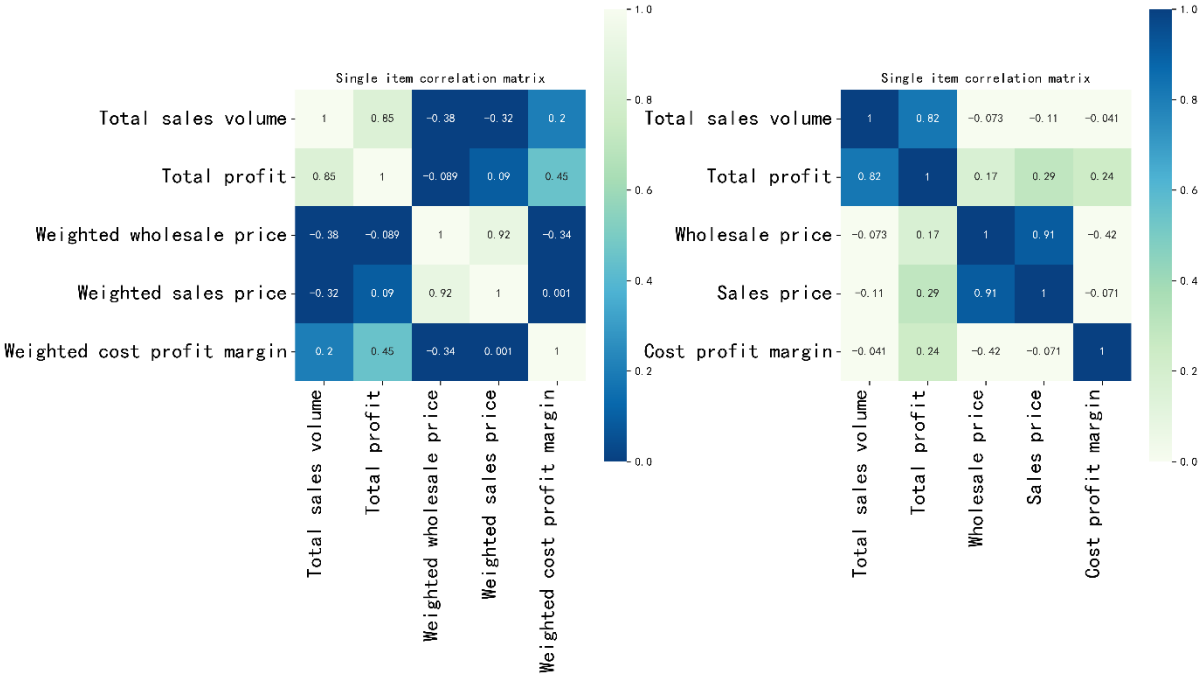
For the wholesale price, sales price and cost-plus margin, this paper considers the sales share of each individual product in the whole category to normalize, so as to get the normalization coefficient of each individual product under the category. As shown in Equation 4.

$$t_i^n = v_i^n / \sum_{i \in n} v_i^n \tag{4}$$

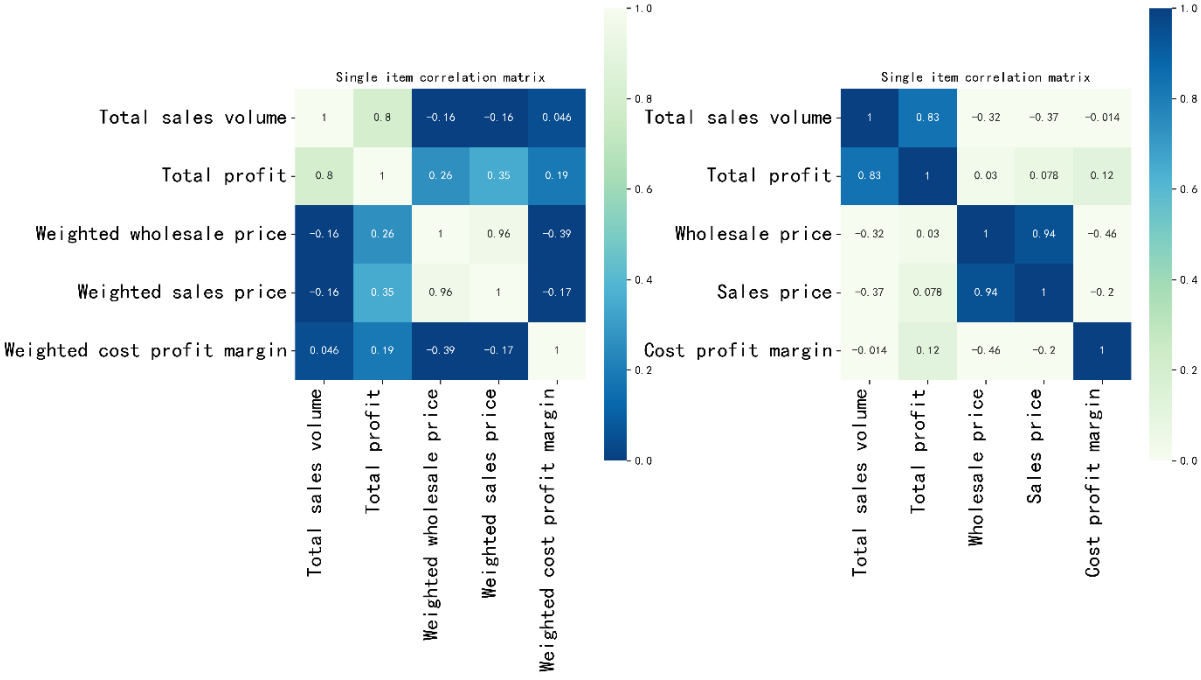
After calculating the normalization factors, the weighted wholesale price, weighted selling price, and weighted cost margin for each category can be calculated, and the results are shown in Equation 5.

$$\begin{cases} p^n = \sum_{i \in n} t_i^n * p_i^n \\ q^n = \sum_{i \in n} t_i^n * q_i^n \\ r^n = \sum_{i \in n} t_i^n * r_i^n \end{cases} \tag{5}$$

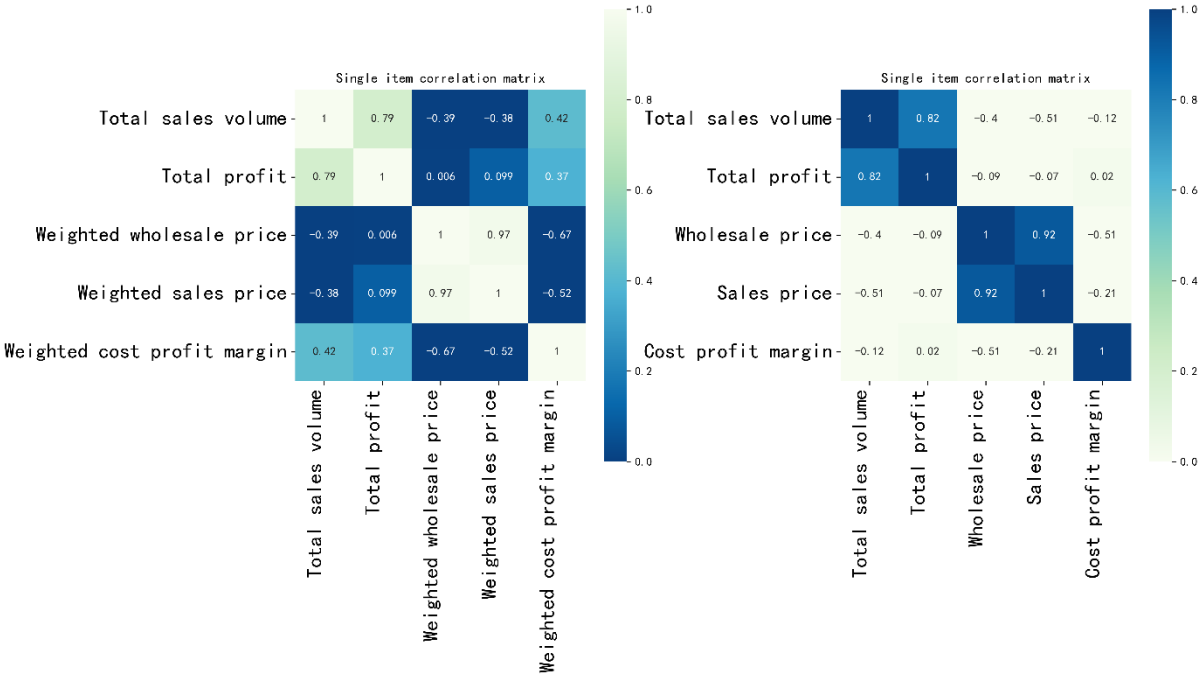
Next, in this paper, heat map visualizations are produced in order, i.e., heat maps of sales versus cost-plus pricing for foliage, edible mushrooms, peppers, cauliflower, aquatic roots and eggplant. The results are shown in Figure 1.



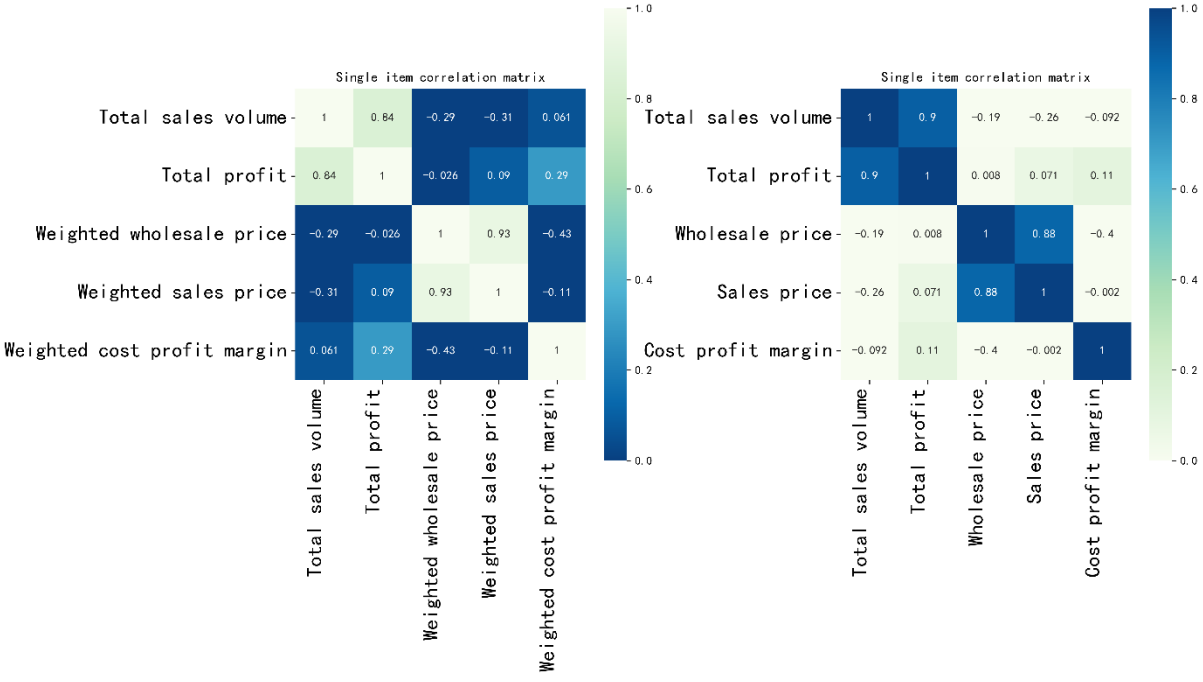
(a) cauliflower



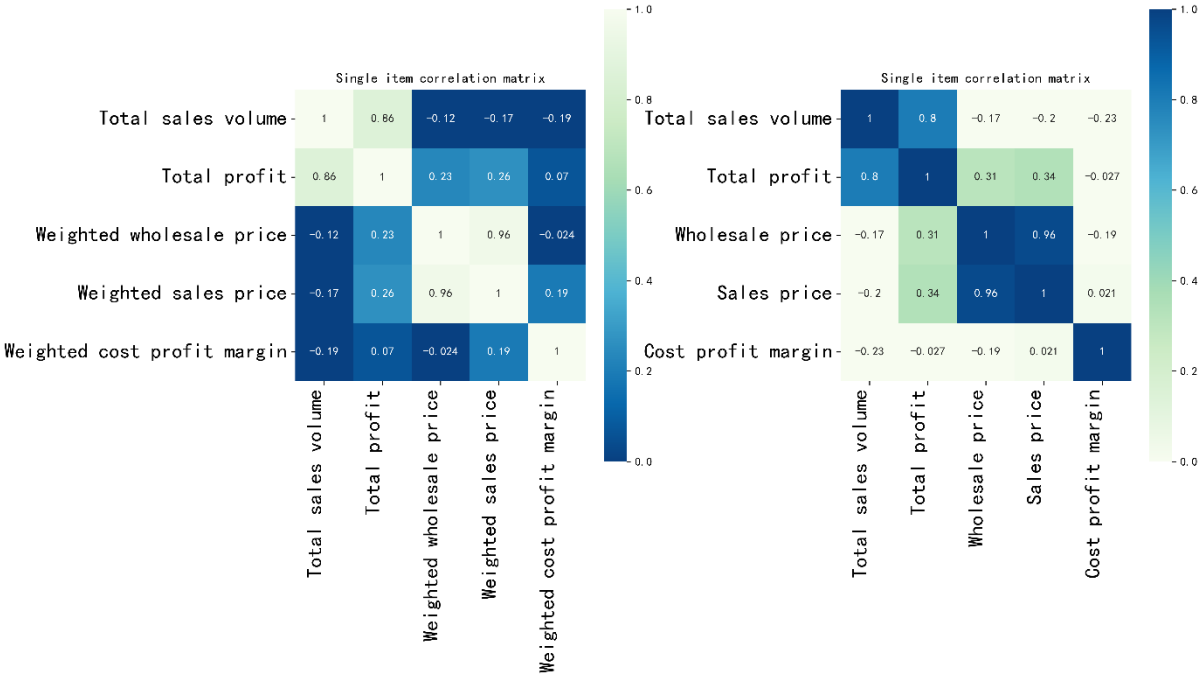
(b)philodendron



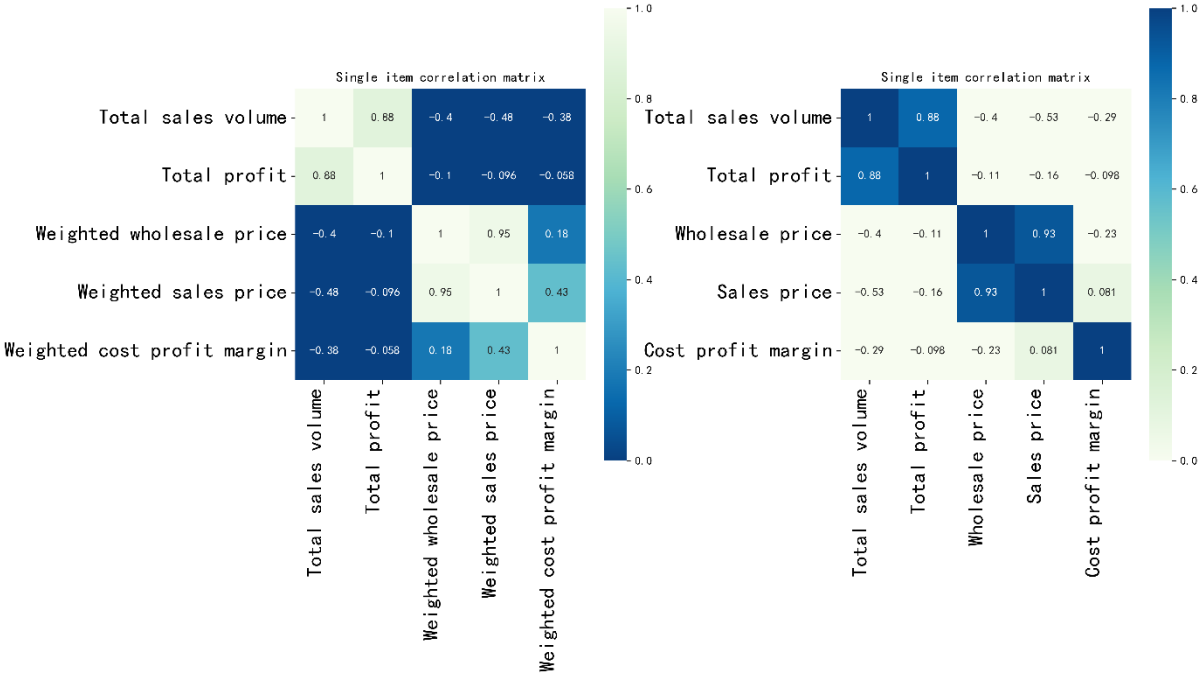
(c)capsicum



(d)eggplant



(e)edible fungi



(f)Aquatic rhizomes

Figure 1. Heat map of sales and cost-plus pricing for foliage, edible mushrooms, peppers, cauliflower, aquatic roots and eggplants.

By looking at the heat map, it can be seen that the correlation between cost margin and total sales is weak, that is to say, the link between cost margin and total sales as well as wholesale price is not significant in either category or individual product correlation analysis. In addition, there is also a relationship between sales price and total profit and cost margin, which is particularly evident in the cauliflower and pepper categories, where there is a significant positive correlation between cost margin and total profit. The higher the cost margin, the relatively higher the total profit. However, this paper argues that this positive correlation is not linear. When the cost margin reaches a certain critical value, the profit will begin to decline and will not continue to grow indefinitely.

The analysis that profits decline despite higher cost margins is not consistent with reality. Normally, higher cost margins would correspond to higher selling prices, but this could lead to a decline in sales.

The reason for this situation may be due to the fact that vegetable products are regarded as necessities and their elasticity of demand is less than 1. This implies that people are less price-sensitive when purchasing such goods and are less susceptible to price changes. Therefore, a moderate increase in cost margins can actually improve profitability.

Building on the above research, further insights were provided into replenishment and pricing strategies from the category to the individual item level. First, in this paper, we screened out the individual items that are available for sale during the period from June 24 to June 30, 2023. For these available items, an Arima model is used to predict their wholesale prices. The Random Forest model is then used to estimate the sales volume of these items. After that, a specific available-for-sale item is selected as an example and described in detail.

2.3. Optimization model

Based on the predicted wholesale prices (predicted by the Arima model), this paper uses different selling prices and combines them with time characteristics such as seasonality and whether it is a business day or not to derive a sales volume forecast. Once the sales volume forecast is obtained, the total profit corresponding to a particular sales price can be calculated. This actually involves an optimization problem whose optimization model is shown in Equation 6.

$$\begin{cases} \max\limits_{\hat{p}^n} \hat{v}^n * (\hat{q}^n - \hat{p}^n) \\ s.t\ \hat{p}^n \le \hat{q}^n \le 3\hat{p}^n \\ e^n \hat{v}^n \le y^n \end{cases} \tag{6}$$

Since this optimization function is a linear optimization problem, in this paper we have chosen to solve this problem by either using a simple linear optimization library or by using a web search.

First of all, some constraints are proposed in this paper, but it is found that these constraints are too idealized, and the actual merchants will have some constraints on the minimum standard of order as well as the number of single items, therefore, this paper will take the minimum standard of displaying the order quantity of 2.5 kilograms, and the number of single items is fixed between 27 and 33 as an example of the specific calculations. In this way, the Pearson correlation analysis can be performed separately for individual items and categories. However, when a fixed number of individual items is required and the minimum display quantity requirement is met, in addition to taking into account factors such as sales and wholesale prices, it is also important to ensure that the number of selected items does not exceed a specific predetermined value and that the minimum display standard is met for each selected item. These additional constraints will further increase the complexity of the problem and require a more detailed optimization approach to solve. As shown in Equation 7.

$$\begin{cases} \max\limits_{\hat{p}^n} \hat{v}^n * (\hat{q}^n - \hat{p}^n) \\ s.t\ \hat{p}^n \le \hat{q}^n \le 3\hat{p}^n \\ e^n \hat{v}^n \le y^n \\ 2.5 \le y^n \end{cases} \tag{7}$$

v_i^n is the random forest predicted sales volume, q_i^n is the pricing strategy, p_i^n is the predicted wholesale price given by the Arima model, e^n is the wear rate for Category N, and y^n is the replenishment strategy. For the number of orders that need to meet the minimum display criteria, with a fixed number of items, it is only necessary to calculate the maximum profit for the day for each

item, and then select the front items in order of profit from highest to lowest as the items to be sold that day. These items will meet the minimum display requirements and have the potential to maximize profit.

3. Result

3.1. ARIMA model fixed order

First of all, considering that the ARIMA model needs to determine the order of the ARIMA model through the autocorrelation and partial autocorrelation functions, and the cost of determining the order by observing the images of multiple categories artificially is high, this paper firstly generates the autocorrelation function and partial autocorrelation function images of a single category. Figure 2 shows the autocorrelation and partial autocorrelation function images of the categories, whose purpose is to determine the order of the ARIMA model. Next, the ADF smoothing test was performed on the individual items, and Table 1 shows the results of the ADF smoothing test, which indicates that the time series of the individual items are smooth and therefore do not require first-order differencing. Next, the y model for the current period was built based on the residuals + lags of the current period. Table 2 represents the final order of the ARIMA model for this category. In this paper, the final order of the Arima model for leafy vegetables was determined: P was chosen to be 1, AR to be 1, and MA to be 0.

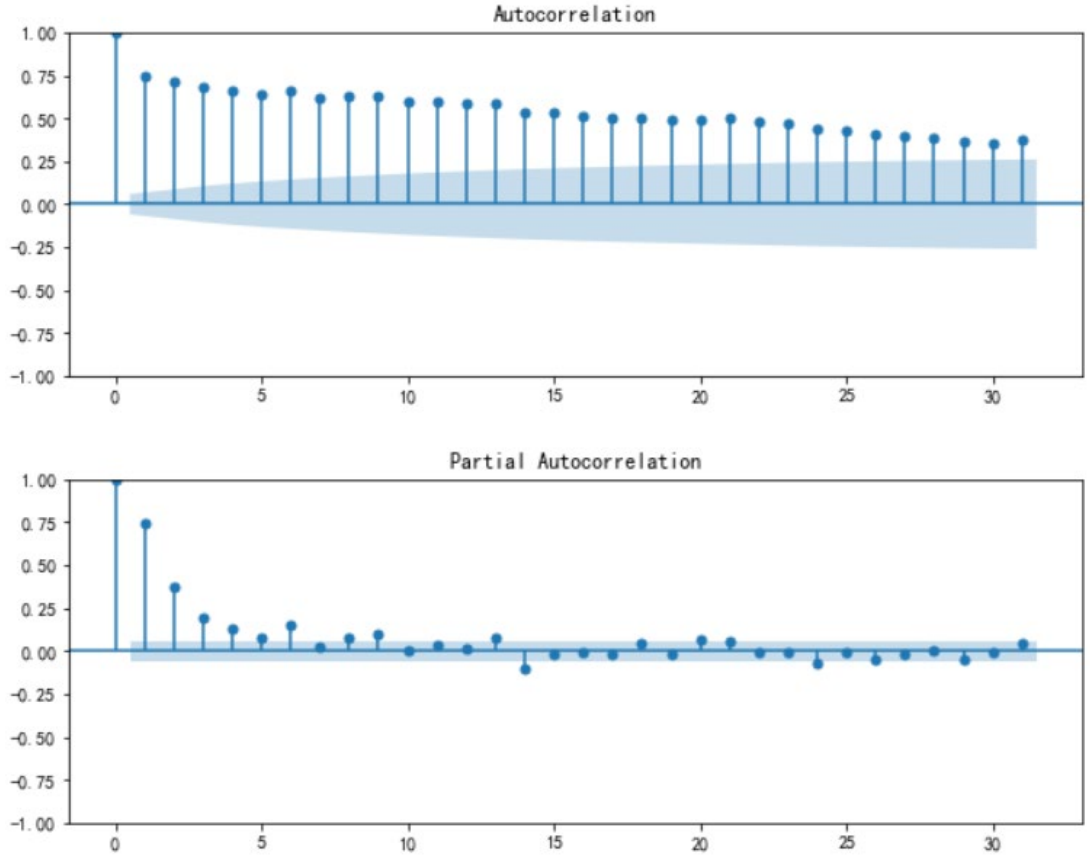


Figure 2. Resulting images of autocorrelation and partial correlation of single products

Table.1. ADF stationarity test

Statistics	*p*-value
-2.93190	0.04176

Table.2. Arima Model Search Results

MA	AR
0	1

3.2. Random Forest Modeling

In this paper, we use Random Forest for multi-feature modeling, which takes into account the wholesale price, selling price, seasonal features and temporal features such as weekdays to jointly predict the sales volume. Firstly, the data of saleable individual items are divided into training and testing sets, then the model is trained on the training set and tested on the testing set, and the corresponding training and testing results are output. Table 3 lists the values of the prescribed random forest model parameter settings. Table 4 shows the results of the training and test sets.

Table.3. Random Forest Parameter Settings

Estimators	Criterion	Max_depth	Min_samples_leaf
100	Gini	None	1

Table.4. Random Forest Test Set Training Set Mae Values

Training set	Test set
0.8725	0.50705

Compared to the category random forest modeling, the random forest model for sellable individual items significantly decreased in terms of MAE scores. This may be due to the larger number of sellable individual items and the difficulty of adapting the parameter configuration of a single random forest model to the characteristics of a variety of different individual items, which leads to a decrease in prediction accuracy.
Overall, this score is still acceptable.

3.3. Optimization model results

Based on the previous analysis, the sales volume and wholesale price can be predicted more accurately. Since this paper has previously constructed an idealized optimization model for the actual problem, which is a linear optimization problem, it is possible to choose either a simple linear optimization library or use the network search method to solve the problem.

By this method, the final predictions for different categories were obtained within seven days. Table 5 shows the final prediction results for different categories.

Table.5. Partial Results Exhibition

Product Code	Sale date	Weighted Wholesale Price	Whether working day	Whether Holiday	Season	Item Code	Replenishment Quantity	Pricing Strategy	Estimated Revenue
1	2023-07-01	7.800	0	1	1	10	33.746	14.500	226.097
2	2023-07-01	3.504	0	1	1	26	54.560	6.955	188.331
3	2023-07-01	6.461	0	1	1	2	39.117	10.491	157.656
4	2023-07-01	4.064	0	1	1	28	32.420	8.096	130.712
5	2023-07-01	3.344	0	1	1	22	40.362	6.216	115.913
6	2023-07-01	15.278	0	1	1	0	7.337	30.516	111.801
7	2023-07-01	4.463	0	1	1	5	23.092	8.894	102.331
8	2023-07-01	18.204	0	1	1	45	5.295	36.406	96.380
9	2023-07-01	6.084	0	1	1	7	26.821	9.425	89.629
10	2023-07-01	3.454	0	1	1	39	20.854	6.881	71.467
11	2023-07-01	2.872	0	1	1	33	23.498	5.708	66.639
12	2023-07-01	3.800	0	1	1	8	16.204	7.600	61.577
13	2023-07-01	7.021	0	1	1	24	7.852	14.032	55.047
14	2023-07-01	10.607	0	1	1	11	4.966	21.211	52.654
15	2023-07-01	7.364	0	1	1	12	7.064	14.646	51.442
16	2023-07-01	2.474	0	1	1	3	20.627	4.911	50.265
17	2023-07-01	7.640	0	1	1	16	7.711	13.860	47.960
18	2023-07-01	2.864	0	1	1	27	16.907	5.696	47.883

4. Conclusion

For pricing and replenishment strategies for vegetable products, this paper uses a cost-plus pricing formula to calculate the cost margin for each individual item per day, and then calculates total sales and total profit based on the given data. Sales for each category are then summarized by weighting individual items according to their importance to the category. The relationship between sales and cost margin for each category is then analyzed. An automated pricing and replenishment strategy is derived for each individual item in each vegetable category for the following week. Since the pricing of each individual item will vary, this paper proposes a random forest model to establish the relationship between price and various other aspects such as sales volume. The Arima model is then used to predict the pricing trend for each category or individual item, and finally an optimization model is built to determine the pricing and replenishment strategy when profits are maximized after considering the wastage of each individual item. This helps fresh produce wholesalers accurately forecast demand and prices for fresh produce, leading to better planning of inventory, pricing and sales strategies, and improved sales efficiency. It also reduces wastage of resources due to inventory backlog or disposal, positively impacting the industry's cost and sustainability.

References

[1] YUE Ruixue,TAN Yanwen.Co-influencing factors and macro determinants of vegetable price fluctuations[J].Journal of South China Agricultural University(Social Science Edition),2019,18(03):79-91.

[2] DING Jiaojiao, WU Jianzhai, KONG Fantao. Analysis of price fluctuation law and transmission mechanism of vegetable production and marketing: Based on ARDL-ECM model[J]. Agricultural Outlook,2019,15(08):11-18.

[3] FANG Wei, YANG Zhenyu, WAN Zhong, et al. Current situation, trend, and countermeasures of consumer demand for major agricultural products in Guangdong Province[J]. Tropical Agricultural Engineering,2016,40(04):58-62.

[4] Sankaran S. Demand forecasting of fresh vegetable product by seasonal ARIMA model[J]. Int. J. of Operational Research,2014,20(3).

[5] PENG Hongxing, ZHENG Kaihang, HUANG Guobin et al.Vegetable price prediction based on BP, LSTM and ARIMA models[J].China Journal of Agricultural Mechanization,2020,41(04):193-199.

[6] CAO Xinyue,HE Chunlin,CUI Mengtian.Urban vegetable price fluctuation law and prediction based on X12-ARIMA and LSTM combination model[J].Journal of Southwest University for Nationalities(Natural Science Edition),2021,47(04):418-425.

[7] Teng Yongping, Huayu.Analysis and short-term prediction of corn price fluctuation in China——Based on X12 seasonal adjustment method, H-P filter method and ARIMA model[J].Journal of Shenyang University of Technology(Social Science Edition),2019,12(03):230-235.

[8] LIU Jie. Application of ARMA model in stock price prediction:Taking Gree Electric Appliances as an example[J].China Management Informatization,2021,24(11):153-155.

[9] Wang Y,Guo Y. Forecasting method of stock market volatility in time series data based on mixed model of ARIMA and XGBoost[J]. China Communications,2020,17(3).

[10] YOU Xianfei. Application of X11-ARIMA model in prediction of hand, foot and mouth disease in China[J].Industry and Technology Forum,2020,19(07):64-66.

[11] LIU C X, ZHAO R B, PANG M Y. A fully automatic segmentation algorithm for CT lung images based on random forest[J]. Medical Physics, 2020, 47(2): 518-529.

[12] PEI Lei. Research on vacuum glass insulation performance detection and vacuum life prediction algorithm[D].Hainan University,2018.