

Sales forecasting and classification model and pricing research of agricultural perishables based on STL-ARIMA

Jian Li[#], Daizheng He[#], Xingshuo Lu^{*,#}

School Of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai, China, 200240

* Corresponding Author Email: Yukino-Yui@sjtu.edu.cn

[#]These authors contributed equally.

Abstract. The prediction of sales volume for agricultural perishable goods is crucial for small-scale merchants to formulate corresponding procurement strategies. Accurate sales volume forecasting helps merchants reduce the risk of excess inventory, establish prudent and appropriate procurement strategies, and provide suggestions for product pricing. To further accurately explore the sales volume patterns of various perishable goods, this study classifies existing data into individual products using the STL (Seasonal-Trend decomposition using Loess) algorithm, revealing the characteristics of each product's sales volume over time. Based on the existing sales volume data, the ARIMA model is used to predict the sales volume of specific categories of goods. A multivariate linear regression model related to pricing, based on existing cost and pricing data, is established to provide merchants with pricing suggestions.

Keywords: Agricultural Perishable Products, Sales Volume, STL Decomposition, Multiple Linear Regression, ARIMA Model.

1. Introduction

Perishable goods are products that deteriorate due to various reasons during storage and cannot maintain their normal use over time, shortening their sellable period [1], Agricultural products (such as green vegetables, food, etc.) are typical examples of these. At the same time, the spoilage of these products generates waste that must be disposed of, resulting in high spoilage costs [2]. For merchants, the ordering situation of perishable goods often involves transaction credit, inflation, storage, and other cost-related factors [3]. Therefore, predicting the cost and sales volume of perishable goods can significantly reduce the costs for merchants and increase profits.

Regarding quantitative models for perishable goods ordering, they can be classified according to the product's usage characteristics into models with a fixed usage period, models with a random usage period, and models with a certain proportion of inventory loss [4]. Taking the random usage period model as an example, Harris's classic Economic Order Quantity (EOQ) model [5] assumes that the goods have an infinite shelf life and is not suitable for perishable goods; on the basis of the EOQ model, Wang Daoping et al [6] improved the EOQ model, proposing a perishable goods EOQ model with time-varying demand rates and a three-parameter Weibull distribution for spoilage rates; Zhang Xingong et al [7] also studied pricing strategies involving price discounts based on this. However, in actual conditions, the changes in spoilage rates under different regional and transportation conditions are difficult to quantify using random variables. At the same time, merchants often purchase from local markets based on consumers requirements of always pursuing the freshest ones [8], so for specific merchants at the supermarket level, the loss caused by spoilage rates is almost negligible.

Therefore, this paper focuses on small-scale merchants, dividing data into individual products and categories, ignoring the impact of spoilage rates, and using STL methods to study the data itself. The ARIMA model is used to predict the sales volume of categories, and a multivariate linear regression equation is established for cost, sales volume, and pricing to provide merchants with references for ordering and pricing.

2. Exploration and Analysis of Agricultural Perishable Goods Cost and Sales Volume Data

2.1. STL Decomposition Analysis of Data Characteristics

The STL decomposition method uses the additive principle to decompose time series data, which can divide the sales volume and time series data of perishable goods into three parts: trend variation (T), seasonal variation (S), and residual (I). Trend variation refers to the long-term trend of data, indicating that the data follows a certain pattern of steady increase, decrease, or remains at a certain level over time; seasonal variation represents regular changes within a year according to the season [9]. The flowchart of STL decomposition is shown in Figure 1. As illustrated, the whole procedure of STL decomposition can be summarized as follows:

Step1: Input the time series Y_v that needs to be decomposed. The initialized value of the trend component ($T_v^{(0)}$) and robustness weight (ρ_v) are set at 0 and 1, respectively

Step2: Enter the inner loop. Each pass of the inner loop consists of a seasonal smoothing that updates the seasonal component, followed by a trend smoothing that updates the trend component. Denote $S_v^{(k)}$ and $T_v^{(k)}$ as the seasonal and trend components at the end of the kth pass. The inner loop contains the following six sub-steps to calculate $S_v^{(k+1)}$ and $T_v^{(k+1)}$:

(1) sub-step1: Detrending of Y_v . A detrended series $Y_v - T_v^{(k)}$ is computed.

(2) sub-step2: LOESS for cycle-subseries with robustness weight ρ_v . A cycle-subseries refers to the subseries of values at one position of the seasonal cycle. Smoothed values of cycle-subseries are calculated at all time positions and the positions just before the first and after the last time position. Note that the neighborhood weight in Eq. (1) should be multiplied by the robustness weight ρ_v . The collection of the smoothed values for all the cycle-subseries forms a temporary series $C_v^{(k+1)}$.

$$v_i(x) = W\left(\frac{|x_i - x|}{\lambda_q(x)}\right) \quad (1)$$

$$\rho_v = B\left(\frac{|R_v|}{6\text{median}|R_v|}\right) \quad (2)$$

$$B(u) = \begin{cases} (1 - u^2)^2, & 0 \leq u < 1 \\ 0, & u > 1 \end{cases} \quad (3)$$

(3) sub-step3: Low-pass filtering of $C_v^{(k+1)}$. The filter consists of two moving average implementations whose lengths are the seasonal period, followed by one moving average implementation of length three and a LOESS implementation. A new time series $L_v^{(k+1)}$ is obtained through the filtering

(4) sub-step4: Detrending of $C_v^{(k+1)}$. The seasonal component $S_v^{(k+1)}$ is computed by $S_v^{(k+1)} = C_v^{(k+1)} - L_v^{(k+1)}$.

(5) sub-step5: Deseasonalizing of Y_v . A deseasonalized series $Y_v - S_v^{(k+1)}$ is computed.

(6) sub-step6: LOESS for $Y_v - S_v^{(k+1)}$ with robustness weight ρ_v . Through the LOESS implementation, the trend component $T_v^{(k+1)}$ is obtained.

Step3: Decide whether to end the current pass of the inner loop. If the inner loop does not reach the preset number of passes, let $T_v^{(k)} = T_v^{(k+1)}$ and return to Step 2; otherwise, exit the inner loop and enter the following outer loop.

Step4: Carry out the outer loop. Firstly, let $T_v^{(k)} = T_v^{(k+1)}$, $S_v = S_v^{(k+1)}$, $R_v = Y_v - S_v - T_v$. Secondly, if the outer loop reaches the preset number, exit the outer loop and output T_v , S_v , and R_v , which are the trend, seasonal, and remainder components, respectively; otherwise, compute the updated robustness weight ρ_v by Eq. (2), in which B is the bisquare weight function expressed as Eq. (3). After acquiring an updated ρ_v , reenter the inner loop, in which the updated ρ_v is used to implement LOESS of Sub-step 2 and 6.

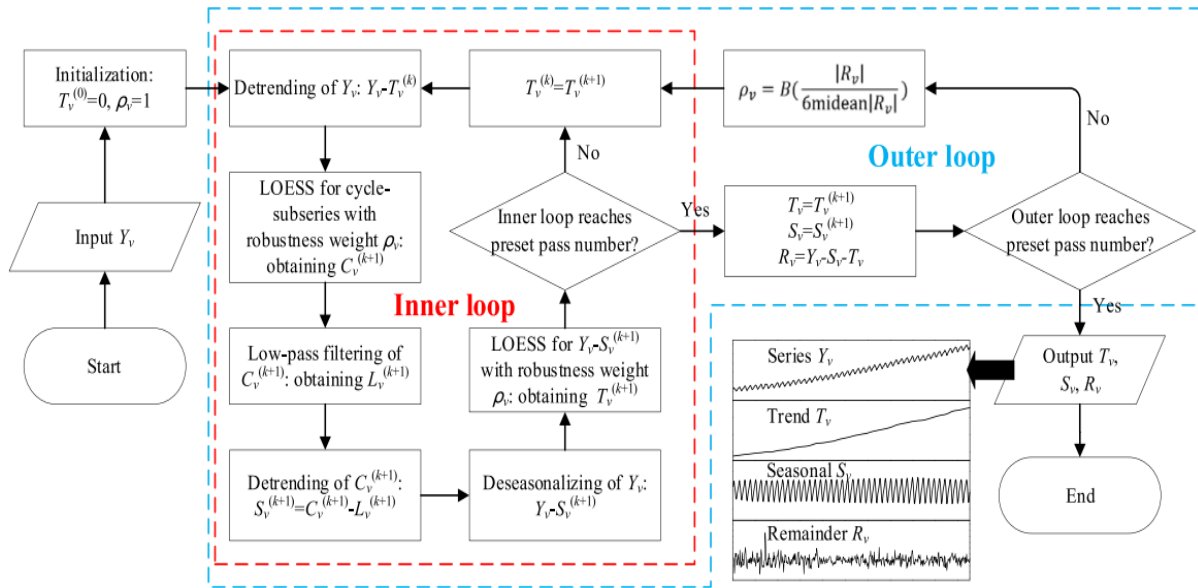


Figure 1. Procedures of STL decomposition.

By using the STL decomposition method to analyze the data of different individual products, the characteristics of the data can be obtained. Archiving this type of data into an existing database can provide references for merchants' choices.

2.2. ARIMA Forecasting of Time Series

The ARIMA model mainly consists of three parts: the autoregressive model (AR), the differencing process (I), and the moving average model (MA) [10]. It assumes that the label value fluctuates around the long-term trend of time, where the trend is composed of historical labels, and the fluctuations are caused by random events over a certain period, and the long-term trend itself may not be stable. Based on this, the hidden time series patterns behind the data can be extracted through the autocorrelation and differencing of the data, and these patterns can be used to predict future data. The mathematical formula is as follows, with a second-order differencing for the differencing part.

$$AR: Y_t = c + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \cdots + \varphi_p Y_{t-p} + \xi_t \quad (4)$$

$$MA: Y_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q} \quad (5)$$

3. Results

3.1. The establishment of STL-ARIMA model

The implementation of the STL decomposition method is based on Eviews software, and the implementation of the ARIMA model is based on SPSSPRO software.

3.2. Analysis of experimental results

Data source: http://www.mcm.edu.cn/html_cn/node/c74d72127066f510a5723a94b5323a26.html.

Using the STL decomposition method, all data can be decomposed. The results of the analysis are shown in Figure 2(Due to space, only part of the presentation is done here). By visually analyzing the number of peaks in the trend, all agricultural perishable goods can be roughly classified into five types: upward trend, downward trend, flat trend, sudden surge, and long-term depression. The characteristics of each type of individual product and the reasons for classification are analyzed as follows:

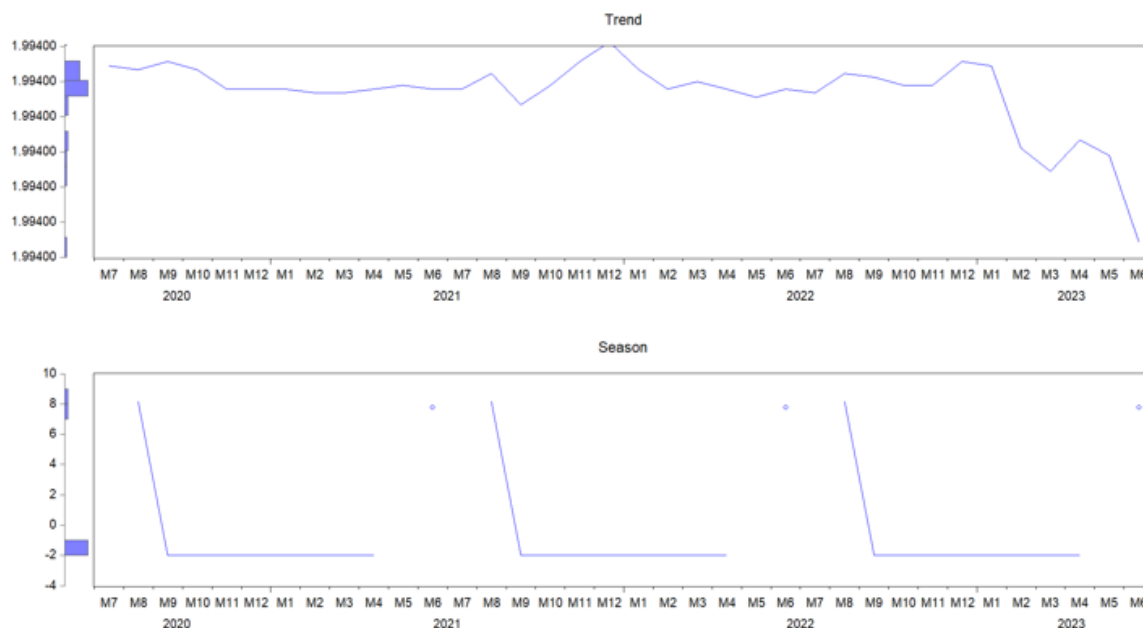
(1) Upward trend: Individual product Water Chestnut, code 102900011001691, shows a general upward trend in sales volume over time in terms of trend variation factors; in terms of seasonal variation factors, the sales volume curve shows peaks during specific periods each year, with summer being the peak season. This type of vegetable is generally popular locally, leading to an increase in replenishment by the supermarket, which in turn causes the sales volume to keep rising.

(2) Downward trend: Individual product Chinese cabbage, code 102900005115793, shows a general downward trend in sales volume over time in terms of trend variation factors; in terms of seasonal variation factors, it reaches a peak during specific periods each year, with summer being the peak season. This type of vegetable is generally unpopular locally, leading to a decrease in replenishment by the supermarket, which in turn causes the sales volume to keep declining.

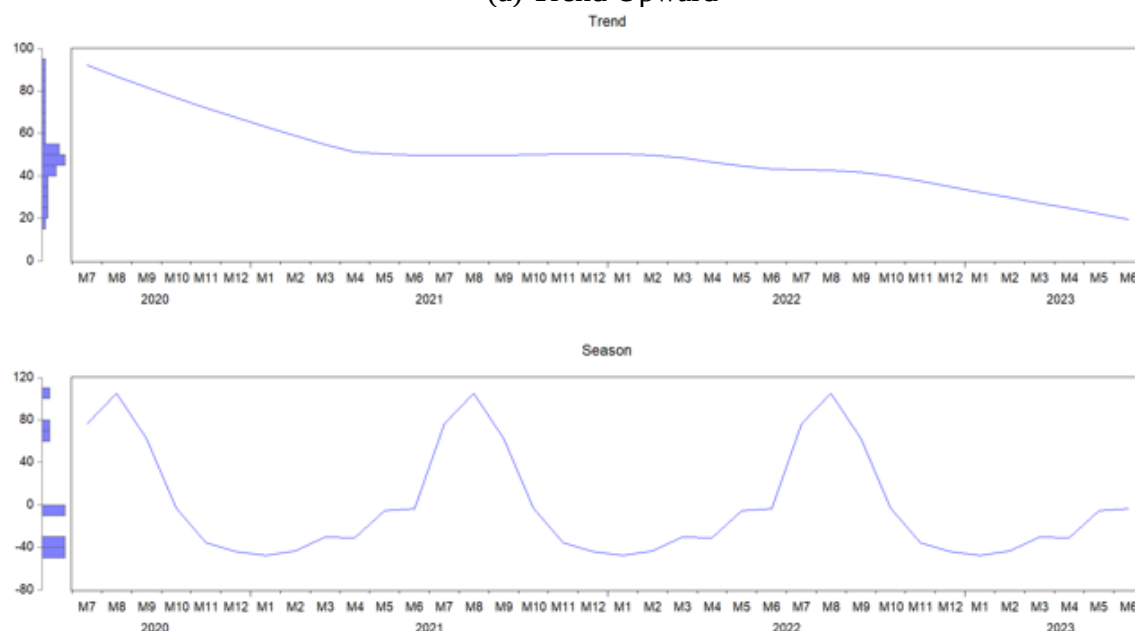
(3) Flat trend: Individual product Amaranth, code 102900005115762, shows a general flat trend in sales volume over time in terms of trend variation factors; in terms of seasonal variation factors, it reaches a peak during specific periods each year, with spring being the peak season. This type of vegetable is generally a staple in the local area, with stable customer demand, and the supermarket's replenishment and sales volume also maintain a certain value.

(4) Sudden Surge: Individual product Local Baby Bok Choy, code 102900005115625, shows no significant trend variation in terms of trend variation factors. However, its sales volume remains at a low level for a long time, with only one or two months of sudden increase in recent years, which does not affect the overall trend. This could be due to sudden events such as online promotion that temporarily boosts sales.

(5) Long-term Depression: Individual product Cordyceps Flower (Box), code 102900011011782, shows a long-term depressed sales volume in terms of trend variation factors, with a general trend that remains unchanged over time, without any seasonal or cyclical variation factors affecting it. However, the sales volume remains at 0 or a low value for a long time. This could be due to the vegetable not aligning with local dietary habits, leading to its long-term lack of popularity in the area.



(a) Trend Upward



(b) Trend Downward

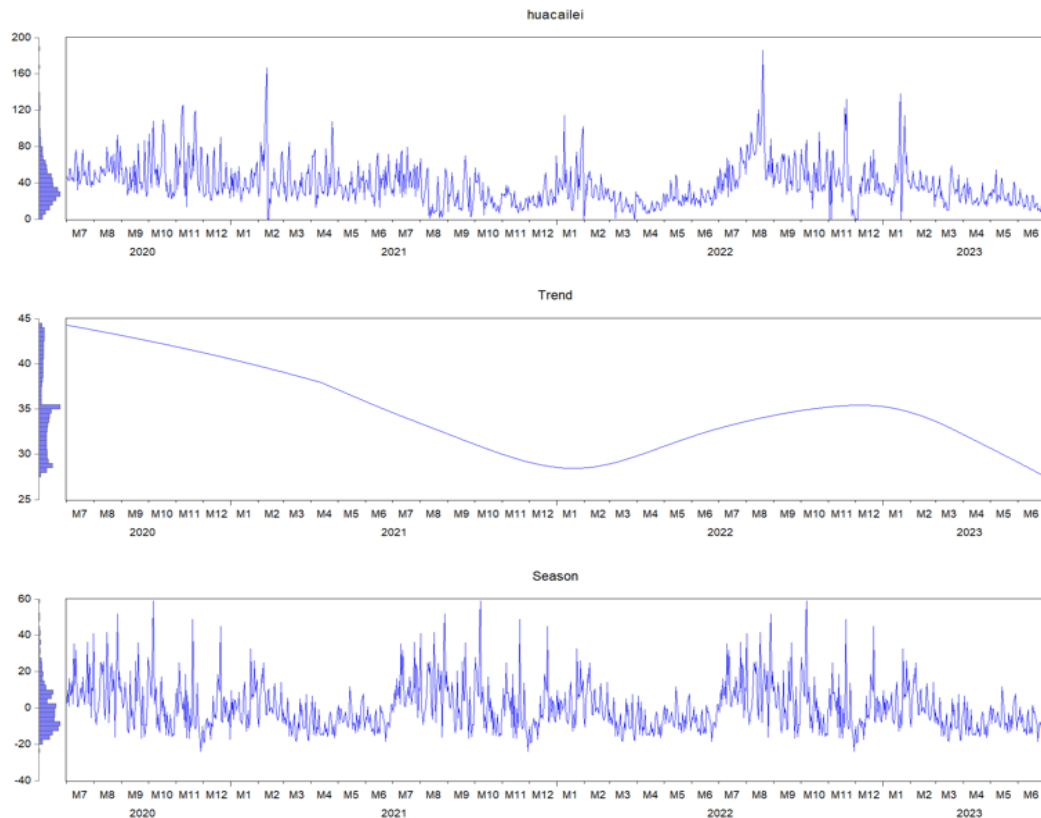
Figure 2. STL Decomposition Results and Classification (Item).

For vegetable categories, similar STL decomposition analysis can be conducted. The results of the analysis are shown in Figure 3(Due to space, only part of the presentation is done here), where six categories are analyzed, and the following results can be obtained:

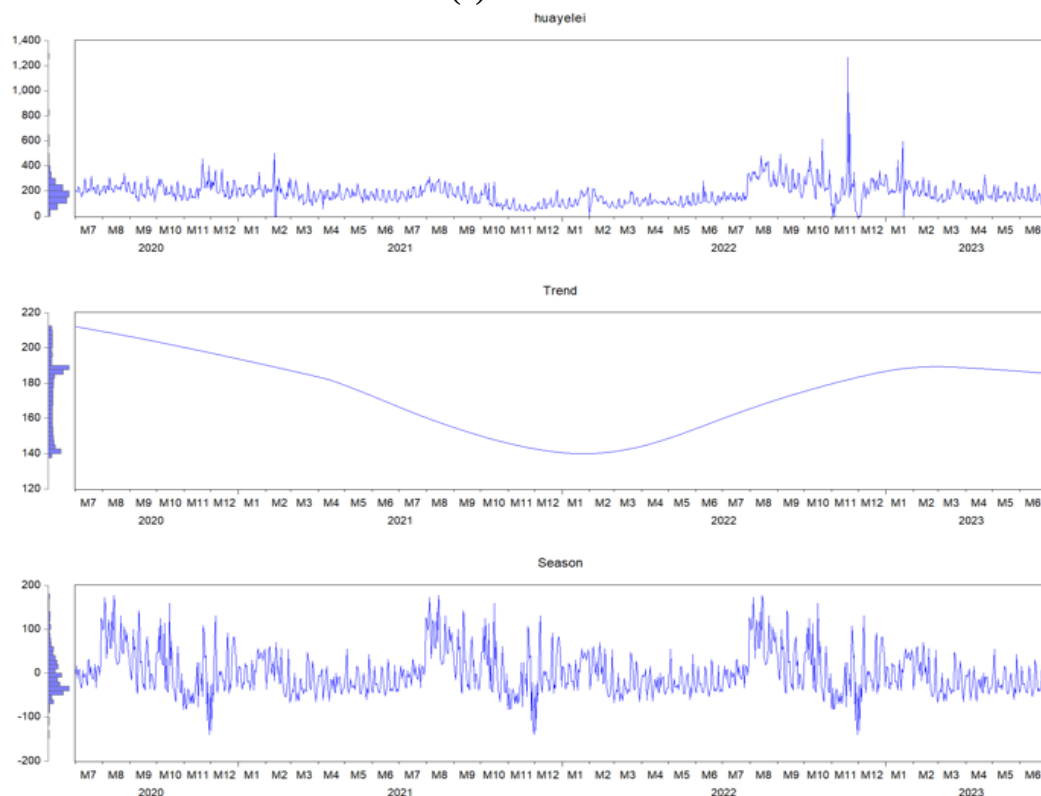
(1) In terms of long-term trends: The sales volumes of various vegetable categories are different. The sales volumes of leafy vegetables, cauliflower, chili peppers, and edible mushrooms first decrease and then increase, with the trough occurring at the end of 2021 and the beginning of 2022. The sales volume of eggplant vegetables shows a long-term downward trend, while the sales volume of aquatic rhizomes is basically on the rise.

(2) In terms of seasonal variation, it is not difficult to find from the figure that each vegetable category has a noticeable peak during a certain period of the year, which is the peak season for the sales of that category of vegetables. Leafy vegetables peak from July to December, cauliflower from September to December, aquatic rhizomes from July to December, eggplant from April to July, chili peppers from January to February, and edible mushrooms from October to the following January.

Merchants can choose the types of vegetables to purchase based on the corresponding trends, reduce the purchase of less favored goods to lower costs, increase the purchase of goods with higher consumer demand to increase income, and adjust orders according to seasonal changes to improve overall income.



(a) Cauliflower



(b) Leaf Greens

Figure 3. STL Decomposition Results and Classification (Category).

Based on the above category analysis, further screening of individual product data within specific major categories can be conducted. For different categories of agricultural products, a multivariate linear regression can be used to obtain the relationship between sales volume (Q), pricing (P), and cost (C), thus assisting in pricing. Since merchants often order goods in large quantities by category, the sales volume, pricing, and cost of specific major category goods can be directly modeled using a multivariate linear regression model. The results of the established multivariate linear regression equations are shown in Table 1.

Table 1. Linear regression equation of total sales, cost and price for each vegetable category.

Category	Equation	R
Cauliflower	$p_t = -0.942 + 0.047 \cdot Y_t + 1.746a_t$	0.137
Leaf Greens	$p_t = -1.899 + 0.009 \cdot Y_t + 2.154a_t$	0.223
Capsicum	$p_t = -1.695 + 0.006 \cdot Y_t + 1.471a_t$	0.289
Solanaceae	$p_t = -0.933 + 0.067 \cdot Y_t + 1.255a_t$	0.331
Edible Mushrooms	$p_t = 7.488 - 0.055 \cdot Y_t + 1.995a_t$	0.103
Aquatic Rhizomes	$p_t = 2.183 - 0.008 \cdot Y_t + 1.512a_t$	0.318

Finally, using the ARIMA model to obtain the predicted sales volume for the next seven days, the results are shown in Table 2. Based on the sales volume of each category and the results of STL decomposition, specific types of individual products that meet the merchant's expectations can be ordered in larger quantities, ultimately helping the merchant to maximize profits.

Table 2. Sales forecast by the ARIMA model for the next seven days.

Date	Cauliflower	Leaf Greens	Capsicum	Solanaceae	Mushroom	Aquatic Rhizomes
1	131.24	21.99	19.96	23.94	81.07	42.85
2	128.28	18.79	20.82	22.00	77.86	42.80
3	136.92	17.84	21.14	21.98	76.20	42.16
4	143.77	17.68	21.44	21.97	74.86	44.67
5	143.18	17.67	21.74	21.95	77.47	45.75
6	138.82	17.66	22.03	21.93	79.00	44.46
7	136.35	17.64	22.32	21.91	79.61	43.83

4. Conclusions

This paper establishes a time series forecasting and classification model for the pricing of agricultural perishable goods based on the STL-ARIMA model. By using existing historical data for prediction simulation and establishing the relationship between cost, pricing, and sales volume using multivariate linear regression, the results show that the STL-ARIMA model can provide relatively accurate predictions while retaining the merchant's choice of goods at specific times.

According to the predicted results, merchants can be divided into two aspects of variety and single item to think about the ordering strategy. For different varieties, according to the overall demand for varieties and pricing changes to develop appropriate average pricing. For a particular variety of items, according to the trend of different appropriate changes in the ordering strategy. For example, order the quantity of stable trend item demand, increase the order quantity of rising trend item, reduce the order quantity of falling trend item, for long-term depressed goods can be appropriately subtracted, for the sudden fire of goods can be appropriately increased a certain amount of order. It's worth noting that the ordering of a single product should meet the demand of forecasting the sales volume of the category and the average pricing.

However, the model itself lacks data on the quantity of goods ordered, and the model is more focused on studying large categories of goods, aiming to provide the simplest, fastest, and most timely information for merchants, without considering the actual spoilage rate. At the same time, both the STL and ARIMA models are highly dependent on the amount of data. The data used in this paper is only 2-3 years, and introducing more data would make the model's prediction more complete. The introduction of unconsidered data (such as spoilage rate) would make the model more complete, but further research is still needed.

References

- [1] GOYAL S K, GIRI B C. Recent trends in modeling of deteriorating inventory [J]. *European Journal of Operational Research*, 2001, 134(1): 1-16.
- [2] DÍAZ R D S, PATERNINA-ARBOLEDA C D, MARTÍNEZ-FLORES J L, et al. Economic order quantity for perishables with decreasing willingness to purchase during their life cycle [J]. *Operations Research Perspectives*, 2020, 7: 100146.
- [3] LUO Z C, LING S N. A review of foreign perishable inventory models and inventory management research [J]. *Logistics Technology*, 2023, 46(07): 149-52+76.
- [4] OU Y F. Research on joint replenishment and cost-sharing strategy for perishable goods with time-varying deterioration rate [D]. 2022.
- [5] HARRIS F W J O R. How Many Parts to Make at Once [J]. 1990, 38(6): 947-50.
- [6] WANG D P, CHEN L, YANG C. Research on the optimization model of perishable goods inventory based on Weibull function [J]. *Industrial Engineering and Management*, 2014, 19(05): 1-6.
- [7] ZHANG X G, MO N. Inventory ordering and pricing strategy for perishable goods based on Weibull function and price discount [J]. *Journal of Chongqing Normal University (Natural Science Edition)*, 2020, 37(04): 1-5.
- [8] CHI S H, ZHANG Q H. Joint decision research on pre-sale inventory and pricing for fresh e-commerce considering consumer freshness preference [J]. *Logistics Engineering and Management*, 2021, 43(08): 116-25.
- [9] HE R, ZHANG L, CHEW A W Z. Modeling and predicting rainfall time series using seasonal-trend decomposition and machine learning [J]. *Knowledge-Based Systems*, 2022, 251: 109125.
- [10] GUI S S, SUN W, XU X F. Analysis of car sales forecasting based on a combined ARIMA and linear regression model [J]. *Computer and Digital Engineering*, 2021, 49(08): 1719-23.