

Statistical Forecasting of Fresh Vegetable Sales Trends: Application of Correlation Analysis and ARIMA Model

Hao Hu^{1,*}, Jiahui Zhang¹, Yuqi Liu²

¹ College of Civil Engineering and Water Resources, Qinghai University, Xining, China, 810016

² School of aerospace engineering, The north university of China, Taiyuan, China, 030051

* Corresponding Author Email: huhao030323@163.com

Abstract. Vegetable commodities have many varieties, short shelf life special purchase times, and other status quo, so the day not sold out of vegetables often can not be sold the next day. Therefore, it is important to accurately predict the demand for vegetable commodities in the coming period and the amount of demand for the replenishment and profitability of vegetable superstores. In this study, Spearman's correlation analysis was applied to investigate the relationship between the daily sales of six major vegetable categories: leafy vegetables, cauliflower vegetables, aquatic root vegetables, eggplant vegetables, pepper vegetables, and edible mushrooms, and the results showed that there was a significant positive correlation between aquatic roots and edible mushrooms which The results show that there is a significant positive correlation between aquatic root vegetables and edible mushrooms, which indicates that consumers who buy aquatic root vegetables tend to buy edible mushrooms as well and that vegetable superstores should take note of this potential relationship to better formulate replenishment strategies; and then the ARIMA model was applied to predict the demand for these six major vegetable categories in the coming week, and the average relative error of the test set was 24.93%, which is a good prediction effect.

Keywords: Descriptive Statistical Analysis, Spearman Correlation Analysis, ARIMA Model, Wavelet Transform.

1. Introduction:

China is a big vegetable-growing country and a big vegetable-eating country. According to a 2019 report by the Chinese Academy of Agricultural Sciences (CASA), excluding production and transportation losses and waste, Chinese people can consume about 0.34 kilograms of vegetables per person per day^[1], and such a high level of demand for vegetables has given rise to a huge number of vegetable superstores. Vegetable supermarkets often sell many types of vegetables, most of which have a relatively short shelf life, and their quality deteriorates with the increase in sales time, if not all are sold on the day, the next day can not be sold, resulting in economic losses. Therefore, it is important to accurately predict the demand for vegetable commodities in the coming period as well as the demand volume for the replenishment and profitability of vegetable superstores.

In 1904, British psychologist Charles Edward Spearman introduced the Spearman correlation analysis method^[2], a versatile tool for assessing monotonic relationships between two variables, widely applicable in psychology, social sciences, biology, and economics. This method, not confined to linear relationships or normal distribution, has been effectively employed in various studies, such as examining the correlation between high school students' English learning enjoyment, independent learning, and performance by Yilun He^[3], and analyzing the relationship between physicochemical factors, flavor substances, and bacterial compositions of cellar mud across different periods by Wu et al^[4].

Similarly, the Autoregressive Integrated Moving Average Model (ARIMA), proposed by Box and Jenkins in 1970^[5], has become pivotal in analyzing and forecasting time series data with cyclical variability across diverse fields like economics, finance, meteorology, and transportation planning. This model, consisting of autoregressive (AR), differencing (I), and moving average (MA) components, is adept at capturing data trends and handling cyclic changes, sudden shifts, or noisy

data. It has been successfully applied in various studies, including Xiao Dan's GDP analysis of Sichuan Province^[6] and Wang Li, Cao Shiyi et al.'s prediction of HFMD infections in Jingzhou City^[7].

In our study, we conducted descriptive statistics and frequency analysis on the daily demand of key vegetable categories for initial data exploration. Subsequently, we utilized Spearman correlation analysis to examine the relationship between these categories' daily demands. Based on these insights, we applied the ARIMA model to forecast future demand. The data for this research was sourced from the China University Student Mathematical Modeling Competition official website (www.mcm.edu.cn). It comprises details of six vegetable categories distributed by a supermarket, along with daily sales data for each item from July 1, 2020, to June 30, 2023. These categories include leafy vegetables (e.g., garland chrysanthemums, Chinese cabbages), cauliflower, aquatic roots and tubers, eggplant vegetables, chili peppers, and edible mushrooms (e.g., monkey-head mushrooms, enoki mushrooms), each identified by a unique numerical code.

2. Data exploration

The data was first processed initially to count the total daily sales for each vegetable category from July 1, 2020, to June 30, 2023, using the date as a reference and considering that the sales volume for that day is the demand for that day.

Then the historical daily sales volume data of each vegetable category were analyzed by descriptive statistics and frequency analysis to explore the data distribution. Here, aquatic root vegetables are taken as an example, and the results of the statistical analysis are shown in Table 1.

Table 1 Data on indicators of aquatic root vegetables

Vegetable category	Aquatic root vegetables
sample size	1085.000
maximum values	296.792
minimum value	0.926
average value	37.402
(statistics) standard deviation	31.357
upper quartile	30.194
skewness	2.504
kurtosis	15.162
variation coefficient	0.838

The results show that the total number of samples is 1085, which corresponds to 1085 days from July 1, 2020 to June 30, 2023, i.e., aquatic root vegetables are sold every day; the average daily sales volume of aquatic root vegetables is 37.402 kg, the minimum daily sales volume is only 0.926 kg, and the maximum daily sales volume is as high as 296.792 kg, the standard deviation of the daily sales volume is 31.357 kg, the data fluctuates greatly; its kurtosis is 12.162 and the skewness is 2.508, the absolute value is less than 3, the data do not conform to normal distribution. The standard deviation of the daily sales volume is 31.357 kg, and the data fluctuates greatly; its kurtosis is 12.162, the absolute value is more than 10 and the skewness is 2.508, the absolute value is less than 3, the data does not conform to the normal distribution.

Further frequency count analysis was performed and the results are shown in Table 2.

Table 2 Results of frequency analysis of aquatic root vegetables

Aquatic root vegetables	interval	frequency	Percentage/%	Cumulative percentage/%
daily sales volume	[0.926,74.892)	986	90.876	90.876
	[74.892,148.859)	89	8.203	99.078
	[148.859,222.825)	6	0.553	99.631
	[222.825,296.792]	4	0.369	100
add up the total	[0.926,296.792]	1085	100	100

The results show that the frequency of daily sales volume in the range of $[0.926, 74.892)$ is 986, accounting for 90.876%; the frequency of daily sales volume in the range of $[74.892, 148.859)$ is 89, accounting for 8.203%; Daily sales volume in $[148.859, 222.825)$ frequency is 6, accounting for 0.553%; daily sales volume in $[222.825, 296.792]$ frequency is 4, accounting for 0.369%.

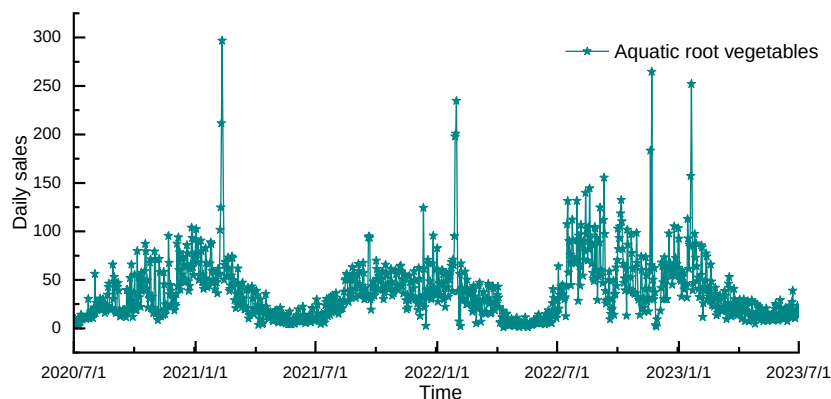


Figure 1 Line graph of daily sales of aquatic root vegetables

Among them, the daily sales volume in the interval $[0.926, 74.892)$ accounted for the highest percentage, as high as 90.876%, and the sum of other intervals accounted for only 9.124%. A line graph of daily sales of aquatic root vegetables is plotted for easy visual analysis, which is shown in Figure 1

It can be seen that the daily sales of aquatic root vegetables roughly show a Corsi distribution and have a cyclical variability with a period of roughly one year.

An effective tool for determining the periodicity of data is the wavelet transform. Sometimes the periodicity of data cannot be detected in the time domain, but if the data is converted to the frequency domain, these periodic characteristics are easily visible^[8].

The key processes are as follows: choose the complex market wavelet function as the wavelet basis function; call the cwt function in the MATLAB library to carry out the continuous wavelet transform; find the wavelet variance, observe the position of the largest wavelet variance x-axis to obtain the main period. A wavelet variance plot of daily sales of aquatic root vegetables was drawn as shown in Figure 2.

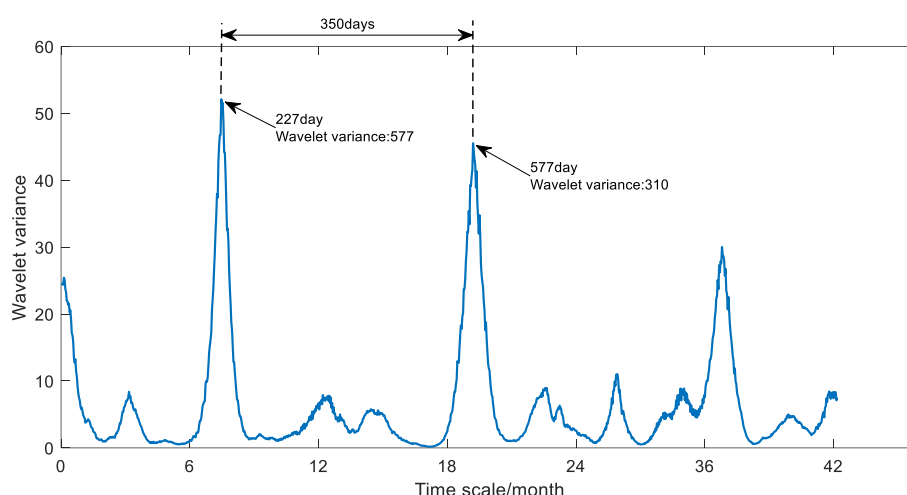


Figure 2 Wavelet variance plot for aquatic root vegetables

According to the wavelet variance plot, the wavelet variance between day 227 and day 577 is more significant, and 350 days (about 1 year) can be identified as a cycle of change in sales of aquatic rhizomes.

3. Vegetable category relationship mining

In the previous phase of work, we obtained the daily sales volume of six vegetable categories for a total of 1085 days from July 1, 2020, to June 30, 2023, and next applied Spearman's correlation analysis to study the relationship between these six vegetable categories through their daily sales volume sequences and performed significance tests.

Spearman correlation analysis calculates the correlation coefficient based on the rank order of the two variables, which refers to the position of the observations in the ordered sequence. The correlation coefficient captures the association between the direction and strength of the trend of change in the variables being analyzed. The correlation coefficient has a value range of $[-1, 1]$, and the closer the absolute value is to 1, the stronger the correlation is between the variables^[9].

The formula for Spearman's correlation coefficient is shown in equation (1).

$$\rho = 1 - (6 \times \sum d^2) / (n \times (n^2 - 1)) \quad (1)$$

where ρ denotes Spearman's correlation coefficient, $\sum d^2$ denotes summing over all ordinal differences squared, and n denotes the number of samples.

The results of the significance test are generally expressed in terms of the P parameter. The smaller its value, the more significant the result is and the more credible the correlation between the two vegetable categories is^[10].

The results of the analysis can be visualized by correlation heatmaps, with ***, **, and * representing the significance levels of $P=0.01$, 0.05 , and 0.1 , respectively.

The results are shown in Figure 3

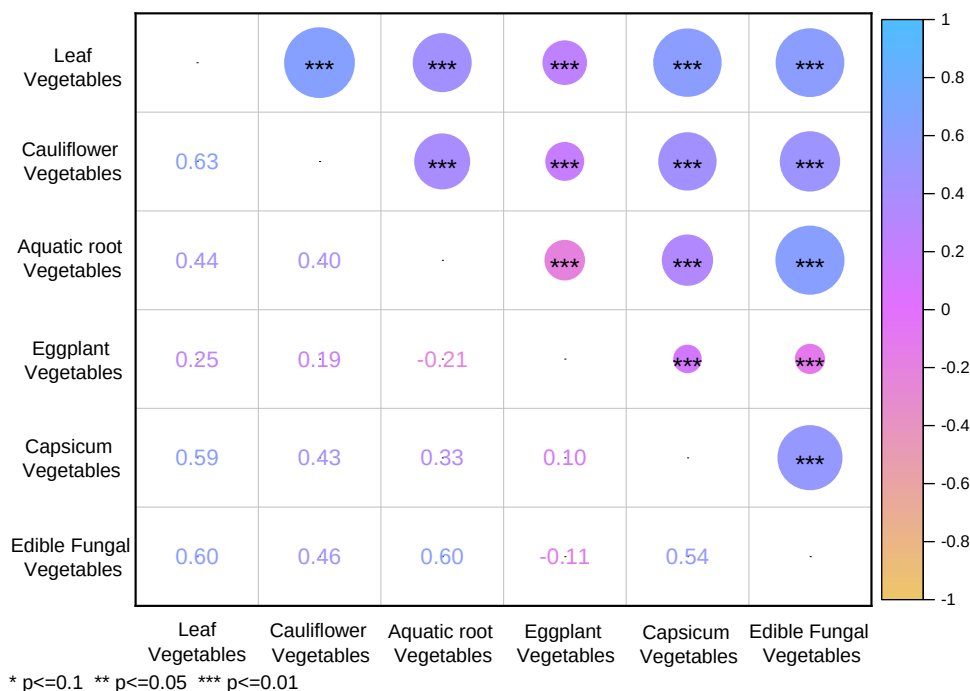


Figure 3 Heat map of Spearman's correlation for each category of vegetables

Analyzing the heat map from a vegetable category correlation study reveals significant correlations among different types of vegetables. For instance, there's a discernible weak negative correlation between eggplant vegetables and both aquatic root vegetables and edible mushrooms. Conversely, eggplants show a weak positive correlation with chili peppers, leafy vegetables, and cauliflower. In contrast, aquatic root vegetables and edible mushrooms exhibit a strong positive correlation, similarly sharing this strong positive relationship with pepper, leafy, and cauliflower vegetables. Edible mushrooms, too, have robust positive correlations with chili peppers and leafy vegetables, alongside a weaker positive link with cauliflower. Additionally, a strong positive correlation exists between

chili and leafy vegetables, as well as a mild positive connection with cauliflower. Notably, leafy and cauliflower vegetables also share a strong positive correlation.

These patterns hint at the associative behaviors consumers exhibit when purchasing vegetables. For example, those buying aquatic root vegetables often prefer edible mushrooms over eggplants, evidenced by the strong positive correlation with the former and weak negative with the latter. Such insights can be invaluable for supermarkets in developing effective replenishment strategies, taking into account these potential consumer purchasing behaviors.

4. ARIMA model to forecast future demand

After obtaining the relationship between each vegetable category, the ARIMA model can be applied to predict its daily sales volume in the next 7 days based on its historical daily sales volume data, and it is considered that the daily sales volume in the next 7 days is its daily demand. Here we take the chili pepper vegetables as an example.

The data of the first 35 months of its historical daily total sales volume sequence is taken as the training set and the data of the last 1 month is divided as the test set. Draw the original time series plot of the training set of daily sales of chili vegetables i.e. Figure 4.

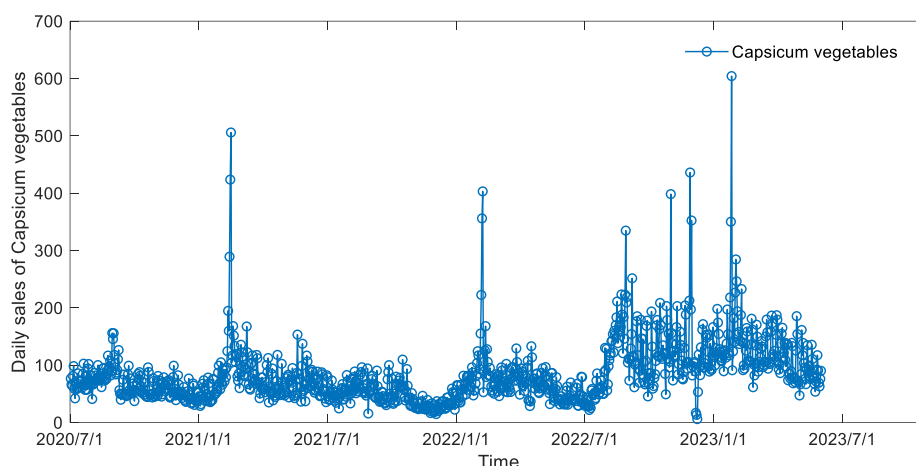


Fig. 4 Original timing diagram of the training set of daily sales volume of pepper vegetables

The time series is determined to be smooth by the Augmented Dickey-Fuller test (unit root test) and therefore does not need to be differentiated^[7].

Based on the correlation characteristics, the model type can be identified by the trailing and truncated nature of the autocorrelation function and the partial correlation function, i.e., using the partial correlation function (PACF), to determine the lag order of the AR model; using the autocorrelation function (ACF), to determine the lag order of the MA model^[11]. The plots of its autocorrelation function and partial correlation function are shown in Figure 5.

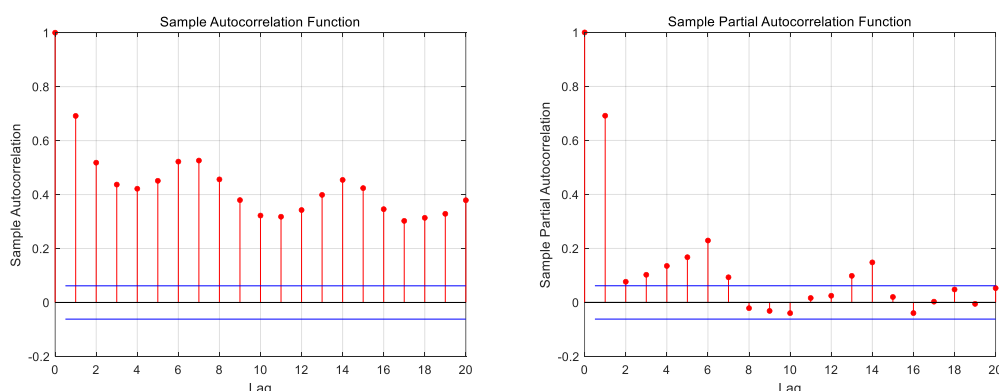


Fig. 5 Plot of autocorrelation and partial correlation function of pepper vegetables

The autocorrelation function and partial correlation function plots are not good for determining the model order, which can be calculated by quantitatively analyzing the *AIC* criterion. Equation (2) is the *AIC* calculation formula in MATLAB.

$$AIC = (-2 \times LLF) + 2 \times \text{ParNum} \quad (2)$$

LLF is the Great Likelihood Ratio and *ParNum* is the number of parameters to be estimated. a smaller *AIC* means a better model.^[11].

The final obtained *AIC* fixed-order thermogram is shown in Figure 6.

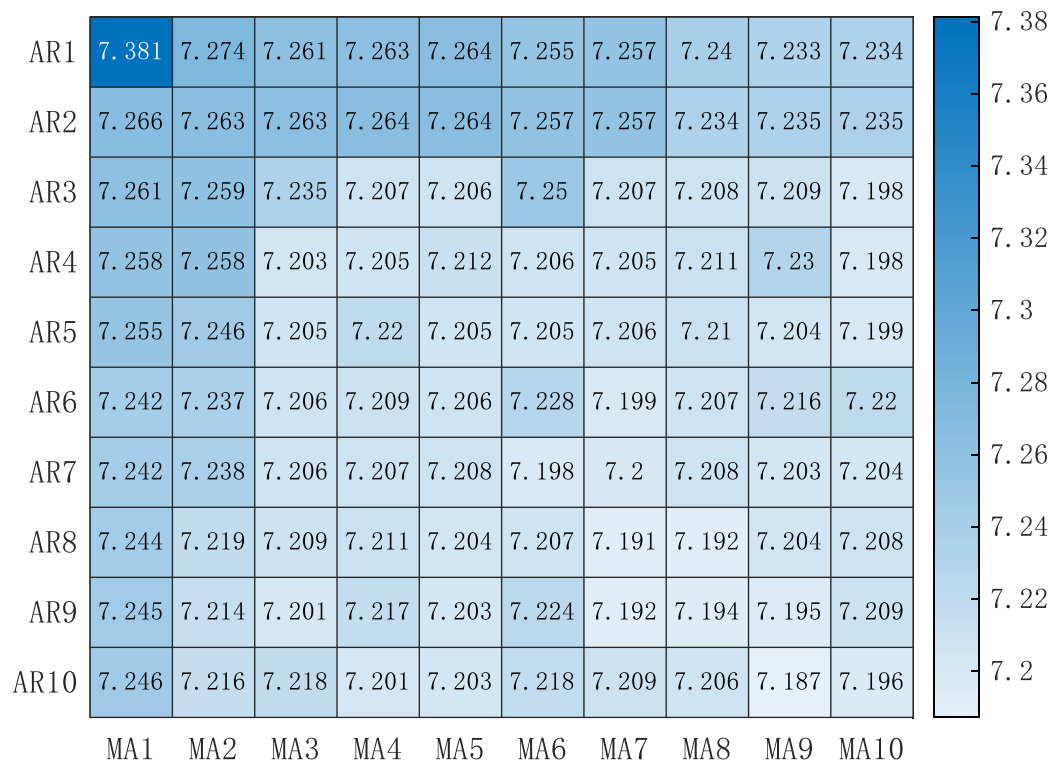


Figure 6 Fixed order thermogram of AIC for chili vegetables

As can be seen from the above figure, the *AIC* is minimized when $p = 10$ and $q = 9$, which identifies the model as the ARIMA(10,9) model. The prediction of the ARIMA(10,9) model on the validation set is shown in Figure 7.

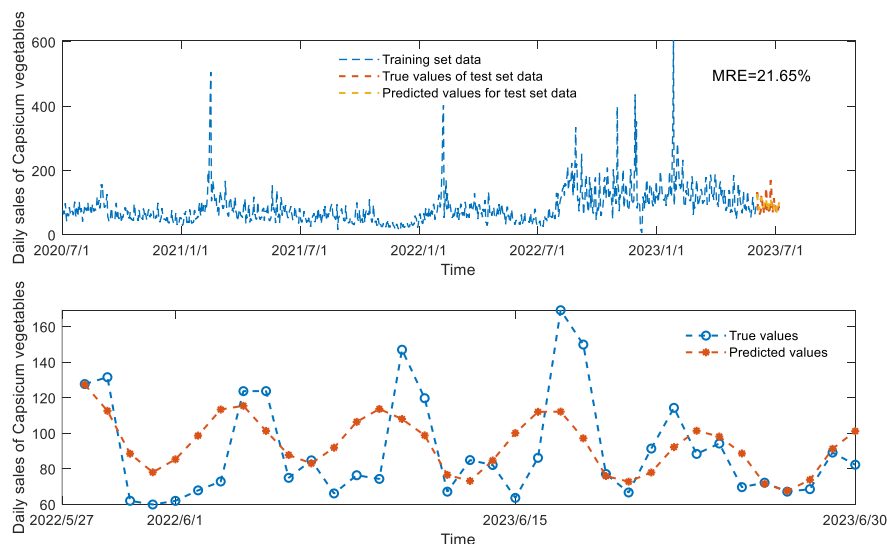


Figure 7 Effectiveness of validation set prediction for pepper vegetables.

The average relative error is 21.56% and the model achieves a high level of accuracy.

Continuing to invoke the ARIMA(10,9) model to forecast the daily demand for July 1-7, 2023, i.e., the coming week, the following results were obtained and are shown in Figure 8.

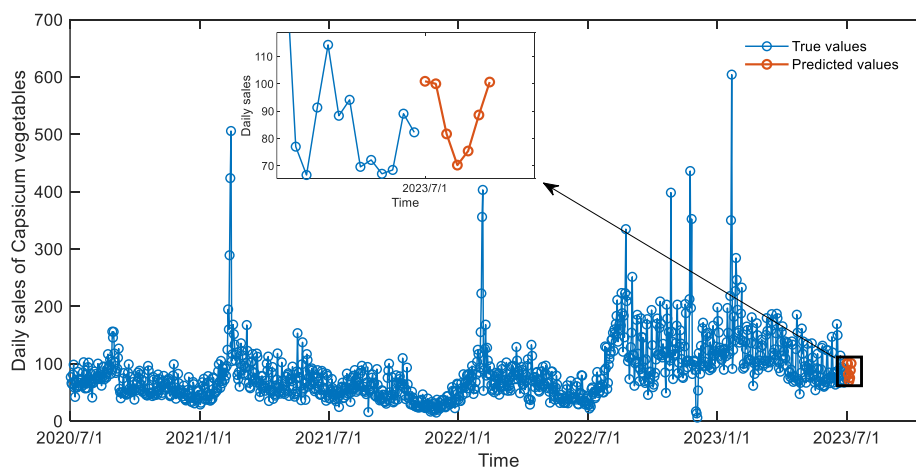


Figure 8 Forecasted results of daily demand for chili vegetables
Specific projected quantities are shown in Table 3.

Table 3 Forecast results of daily demand for chili vegetables

Date	Daily demand/Kg
2023/7/1	100.96
2023/7/2	100.05
2023/7/3	81.71
2023/7/4	70.29
2023/7/5	75.45
2023/7/6	88.63
2023/7/7	100.68

The same forecast for the remaining five vegetable categories yields the projected values of their daily demand for July 1-7, 2023, and the detailed results are shown in Table 4.

Table 4 Predicted results for the remaining five vegetable categories

Date	Leafy vegetables	Cauliflower vegetables	Aquatic root vegetables	Eggplant vegetables	Edible Fungal Vegetables
2023/7/1	179.19	30.62	17.39	25.58	93.98
2023/7/2	184.68	31.10	13.36	30.31	82.34
2023/7/3	163.96	22.24	8.23	23.38	63.85
2023/7/4	162.40	21.32	10.80	20.43	66.87
2023/7/5	150.84	21.70	10.68	18.31	56.10
2023/7/6	126.30	19.28	9.69	17.33	51.79
2023/7/7	148.03	22.43	15.15	21.01	70.58

The average relative error for a total of 6 predictions is 24.93%, the model has high accuracy and validity.

5. Conclusions

This paper provides a research idea and framework for the reference of vegetable category merchandising strategy. Firstly, Spearman correlation analysis is applied to study the relationship between the sales volume of each major vegetable category, and for the vegetable categories presenting significant correlation, they can be linked for more in-depth consideration when

wholesaling; then ARIMA model is applied to predict the demand for each category of vegetables in the coming week, and according to its demand, the wholesale quantities of each category of vegetables are determined more scientifically to minimize the loss of the superstore and to increase the revenue. increase revenue.

The average relative error of the ARIMA model for the test set in a total of six predictions is: 24.93%, the model achieves a high level of accuracy, proving the feasibility of the method.

References

- [1] Institute of Resources and Agricultural Zoning, Chinese Academy of Agricultural Sciences. China Agriculture Green Development Report 2019 [J]. 2020.
- [2] Chen Qianqian, Hu Bo, Jia Lei. Research and practice of VoLTE single-pass analysis optimization system based on Spearman algorithm, F, 2020 [C].
- [3] He Yilun. A study of the correlation between high school students' English learning pleasure, independent learning, and English achievement [D], 2023.
- [4] WU Langtao, DING Wenjun, XIE Zhuan, et al. Correlation of physicochemical factors, flavor substances, and bacterial composition of cellar soil at different periods [J]. Microbiology Bulletin, 2022, 49(03): 1030-47.
- [5] SUN Changshuai, WANG Jiwen, YU Tianwen, et al. Applicability of ARIMA method for predicting settlement of blown-in soft foundation [J]. Computer Aided Engineering, 2023, 32(01): 69-73.
- [6] Xiao Dan. GDP Analysis and Forecasting of Sichuan Province Based on ARIMA Model [J]. Productivity Research, 2023, (10): 62-6.
- [7] Wang L, Cao S, Song D, et al. Application of ARIMA model in predicting hand, foot, and mouth disease epidemic in Jingzhou [J]. Chinese Journal of Social Medicine, 2023, 40(05): 623-6.
- [8] Lou Zesheng, Zhang Gaobo, Jia Xiangyu, et al. Characterization of typhoon weather GNSS ZTD period based on FFT, and wavelet transform [J]. Geodesy and Geodynamics, 2023, 43(10): 1032-8.
- [9] Li Jiayi. Applicability of Spearman's rank correlation coefficient method based on the integrated water quality pollution index to analyze the automatic water quality monitoring in Jinjiang River Basin [J]. Journal of Environmental Science, 2022, 41(04): 88-92.
- [10] SONG Ye Ye, HUANG Yuan Qiang, JIANG Yi Qing, et al. Statistical significance in data analysis of ultraviolet measurements [J]. China Lighting Appliance, 2023, (12): 16-20.
- [11] JIA Rubin, GAO Jinfeng. Time series prediction method of dissolved gas content in transformer oil based on ARIMA model [J]. Journal of Zhengzhou University (Engineering Edition), 2020, 41(02): 67-72.