Information Systems and Economics (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/infse.2023.041015
ISSN 2523-6407 Vol. 4 Num. 10

Statistical Analysis of Fresh Produce Retail Data in Convenience Stores and Optimization of Pricing Strategies

Hanzhi Xu

School of Mathematics Science, Shanxi University, Taiyuan, Shanxi, 237016, China

Keywords: Multiple Linear Regression, ARIMA, Simulated Annealing Algorithms

Abstract: The sales volume of various supermarket categories is influenced by market and seasonal factors. In the short term, demand for different categories of vegetables fluctuates. Unreasonable purchasing and pricing strategies can inevitably lead to losses. To maximize supermarket profits, a comprehensive analysis of automatic replenishment and pricing strategies for vegetable products is of great significance for fresh food supermarkets. We utilize multiple linear regression to model the sales volume of each vegetable category, coupled with ARIMA-derived wholesale prices for each category in the next 7 days for planning. Finally, we employ simulated annealing to determine the optimal replenishment and pricing strategies, aiming to maximize revenue in the future 7 days.

1. Introduction

In recent years, with the rise of new sales models such as supermarkets and the increasing demand for high-quality vegetables. The number of vegetables that supermarkets need every day is generally increasing. However, for the agricultural products sold by supermarkets: on the one hand, their quality is objectively affected by time and has the characteristics of short shelf life, so unsalable will inevitably lead to the decline of their quality and reduce their competitiveness. On the other hand, due to the variety of vegetables purchased by supermarkets every day and the existence of obvious seasonal dishes, blindly purchasing will make the types and commodities of the vegetables imported a certain degree of disconnection from the market, and ultimately make supermarkets lose money. Therefore, it is very important to give the optimal replenishment and pricing strategy to maximize the profit of supermarkets.

The multiple linear regression, ARIMA, and simulated annealing algorithms used in this paper have a high frequency of use and excellent solution effects in various fields.

Multiple linear regression, as an improved model of linear regression, was first used in the field of economics, and was initially used to solve the situation that "one variable is affected by multiple variables". Nowadays, multiple linear regression has been widely used in economics, industrial technology, computer science, environmental science, and other fields. Han et al applied this method to the study of pixel point matching of disparity jump in-depth images[1]. Liu et al applied this method to the study of pollution source analysis in Chang Tan Reservoir, Zhejiang Province[2].

ARIMA was first proposed by Box, an American statistician, and Jenkins, a British statistician in

the early 1970s. It is a time series analysis model that adds different operations based on the ARMA model. The autoregressive model (AR), the moving average model (MA), and the difference method are combined to obtain the difference autoregressive moving average model. Because it is more comprehensive and scientific than other basic prediction models, the model has been used in the prediction of time series. Zhang et al applied this method to predict the runoff of the lower Yellow River[3].Yu et al applied this method to predict the water level of the middle Yangtze River[4] .

The earliest idea of the simulated annealing model (SA) was developed by N. Annealing. Proposed by Metropolis et al. In 1983, S. Kirkpatrick et al. successfully introduced annealing ideas into the field of combinatorial optimization. It is a stochastic optimization algorithm based on the Monte-Carlo iterative solution strategy, and its starting point is the similarity between the annealing process of solid matter in physics and the general combinatorial optimization problem. Starting from a higher initial temperature, with the continuous decline of the temperature parameter, the global optimal solution of the objective function is randomly found in the solution space combined with the probability jump characteristic, that is, the local optimal solution can jump out of the probability and eventually tend to the global optimal solution. In recent years, it has been used in VLSI optimal design, image processing, combinatorial optimization problems, production scheduling, control engineering, machine learning, neural networks, signal processing, and other fields. Wen et al. Applied this model to estimate the size and density composition of the lunar nucleus[5]. Zhou et al. applied this model to the study magnetotelluric data inversion considering prior information[6] .

In this work, the sales volume and daily pricing data of each category from July 2020 to June 30, 2023 are given. To explore the relationship between the sales volume and the pricing of the supermarket and give the optimal replenishment and pricing strategy, we analyze the problem effectively based on a series of methods such as fitting, forecasting, and planning. Then we use ARIMA to predict the wholesale price of each category of vegetables in the next seven days. On this basis, we combine the constraints to construct a nonlinear programming model and use a simulated annealing algorithm to solve the planning. According to the results of the solution, we formulate the optimization strategy of each category to maximize the income of the supermarket, which can provide some ideas for the actual sales of the supermarket. It can be used for reference.

2. Correlation Analysis of Sales of Vegetables

Correlation analysis should be used to examine the correlation of sales of vegetables, mainly including Pearson correlation analysis and Spearman correlation analysis. The former is used under the condition that the data meets normal distribution; otherwise, Spearman correlation analysis should be used. In the normality test of all vegetables, it was found that only the data of chilies was inconsistent with normal distribution. For the data that mostly follows normal distribution but partially does not, the data can be converted to meet normal distribution before correlation analysis. Based on this, square root conversion was performed specifically to all vegetables, and then Pearson correlation analysis was performed pairwise for the data that followed normal distribution after conversion. The correlation analysis results were visualized in a thermodynamic chart, as shown in Fig. 1.

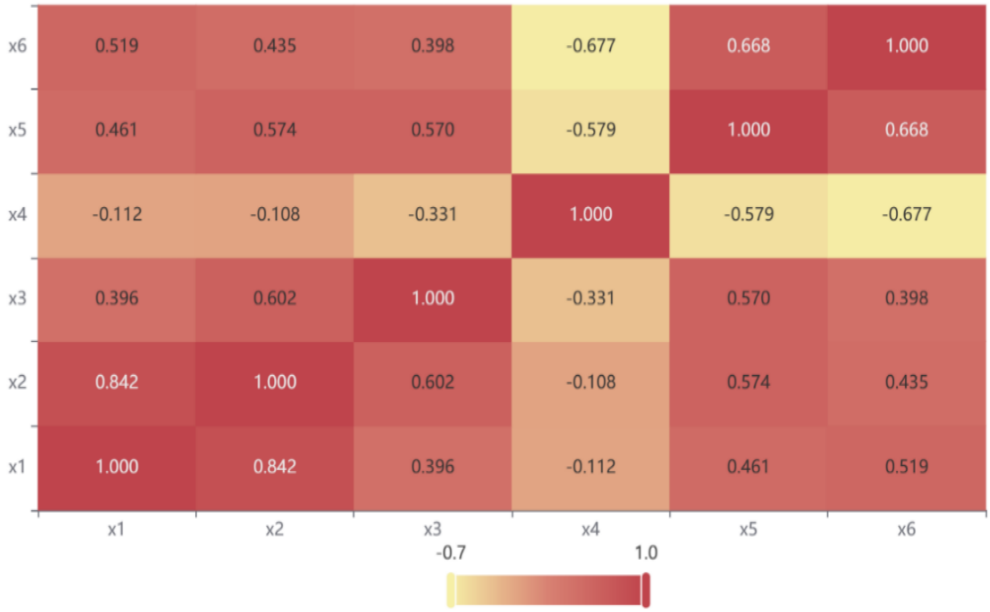

Figure 1: Heat map of correlation analysis of each category.

Where, x1, x2, x3, x4, x5, and x6 respectively represent the unit price of cauliflowers, flower leaves, chilies, eggplants, edible fungi, and aquatic root vegetables.

According to the correlation size, the correlation degree was divided into the following cases, as shown in Table 1.

Table 1: Correlation strength of sales volume data combination for 6 categories

Strength of correlation	Category combination
Strong positive correlation(0.7,1)	Cauliflower-Flower leaves
Moderate positive correlation(0.5,0.7)	Edible fungi-Flower leaves, Chilies-Flower leaves, Aquatic root vegetables-Cauliflower, Edible fungi-Aquatic root vegetables, Edible fungi- Chilies
Moderate negative correlation(-0.7,-0.5)	Eggplants-Aquatic root vegetables, Edible fungi-Eggplants
Strong negative correlation(-1,-0.7)	Nothing

3. Multi-category pricing relationship based on multiple linear regression

3.1 Model non multicollinearity test

Before multiple linear regression, it is necessary to exclude the influence of multicollinearity, which is reflected by the value of the variance inflation factor (VIF). When there is no linear relationship in the variable index n involved in multiple linear regression, VIF$\in$[0,10]. After the calculation of the six sets of data in the above table, the VIF results were all in [0,10), so the data was not significantly collinear, and multiple linear regression analysis could be performed.

3.2 Multiple Linear Regression Results

The correlation analysis only illustrates the relationship between the sales of vegetables, without

considering the influencing mechanism of pricing on the sales of vegetables[7]. The sales of any vegetable are objectively affected by the pricing of all vegetables, so quantitative fitting was performed on the sales of vegetables through regression analysis, as shown in Table 2.

Table 2: Fitting results of multiple linear regression for each category

Category	Multiple regression function
Cauliflower	$f_1 = -1.703x_1 - 1.367x_2 + 0.116x_3 - 0.910x_4 + 0.744x_5 - 0.407x_6 + 66.535$
Flower leaves	$f_2 = 5.259x_1 + -17.267x_2 - 1.354x_3 - 4.623x_4 + 1.020x_5 - 0.001x_6 + 273.871$
Eggplants	$f_3 = 4.212x_1 - 9.874x_2 - 3.553x_3 + 3.979x_4 - 1.540x_5 + 0.812x_6 + 104.544$
Chilies	$f_4 = 0.138x_1 + 1.542x_2 + 0.060x_3 + 0.245x_4 + 0.027x_5\mathrm{k} + 2.017x_6 - 11.064$
Edible fungi	$f_5 = 1.567x_1 - 8.622x_2 + 2.412x_3 + -4.499x_5 - 1.027x_6 + 130.831$
Aquatic root vegetables	$f_6 = 0.413x_1 - 4.952x_2 + 1.192x_3 + 0.477x_4 - 0.277x_5 - 2.349x_6 + 70.471$

Where, x1, x2, x3, x4, x5, and x6 respectively represent the unit price of cauliflowers, flower leaves, chilies, eggplants, edible fungi, and aquatic root vegetables.

3.3 Analysis of experimental results

If the regression of the model is rational, its eigenvalue should be 0 or so, and the eigenvalue of the resulting regression equation should be $10^{\wedge -2}$ at dimension ≥1. This indicates that the model is rational, and the linear regression of multi-category pricing is feasible.

4. Forecast of supermarket wholesale prices in the next 7 days

It is expected that the historical data can be used continuously based on the objective development of things to infer the development trend through statistical analysis. Therefore, ARIMA was used for predictive analysis. See the brief introduction to the principle of the model in the following part.

ARIMA consists of AR, I, and MA, and its final mathematical model is as follows:

#1

Where, y_t is the current value, μ is a constant term, p in order, r_i is autocorrelation coefficient, ε_t is an error, and meanwhile ε_t should follow a normal distribution.

The basis of this formula is that AR and MA models can be applied directly on the assumption that the time series being processed is stable. If the time series is unstable, it is necessary to consider part I of the ARIMA model, namely differential processing.

ARIMA model was used to forecast the wholesale price of each category from July 1 to July 7, 2023. See Table 3 for the parameters (p, d, q) of the model.

Table 3: ARIMA model by category (p, d, q)

Category	(p, d, q)
Cauliflower	(0,1,1)
Flower leaves	(0,1,1)
Eggplants	(2,0,3)
Chilies	(0,1,4)
Edible fungi	(1,1,2)
Aquatic root vegetables	(3,0,1)

ARIMA model was used to forecast the wholesale price of each category from July 1 to July 7, 2023, as shown in Table 4.

Table 4: Wholesale Prices by category from 1 to 7 July 2023 (¥/kg)

Category	1st day	2nd day	3rd day	4th day	5th day	6th day	7th day
Cauliflower	7.897	7.897	7.897	7.897	7.896	7.896	7.896
Flower leaves	3.246	3.245	3.244	3.243	3.241	3.240	3.239
Eggplants	3.573	3.529	3.544	3.569	3.567	3.565	3.563
Chilies	4.673	4.670	4.666	4.662	4.659	4.654	4.650
Edible fungi	4.429	4.338	4.279	4.242	4.217	4.200	4.188
Aquatic root vegetables	17.307	17.329	17.407	17.413	17.412	17.403	17.393

5. Construction of nonlinear programming model

Nonlinear programming was used to optimize the replenishment and pricing strategy of the supermarket in the next 7 days. To solve the problem of the maximum return of supermarket, nonlinear programming was built and a simulated annealing algorithm was used to get the solution. The decision variables, parameters, and constraints of the model were defined as follows:

Decision variables:

x_{it} : pricing of each category, unit: ¥/kg, $i \in \{1,2,...,6\}$ $t \in \{1,2,...,7\}$;

Model parameter:

k_i : loss rate of each category, $i \in \{1,2,...,6\}$;

M_i : maximum pricing of each category, unit: ¥, $i \in \{1,2,...,6\}$;

D_i : minimum pricing of each category, unit: ¥, $i \in \{1,2,...,6\}$;

a_{it} : wholesale price of each category, unit: ¥/kg, $i \in \{1,2,...,6\}$ $t \in \{1,2,...,7\}$;

$f_i(x_{it})$: sales of each category, unit: kg, $i \in \{1,2,...,6\}$, $t \in \{1,2,...,7\}$;

Constraints:

To prevent the programming from extreme cases, it was required that the pricing of each category for all days should not be more than the maximum record and less than the minimum record.

#2

Objective function:

#3

6. Using the simulated annealing algorithm to solve the model

Since the objective function was nonlinear, a heuristic algorithm was required to solve the model, and because the calculation process required a fast local optimal solution, a simulated annealing algorithm was selected.

Based on the wholesale prices of each category predicted by the ARIMA model, we solve the objective function, calculate the above model using simulated annealing, select 10,000 iterations, make the profit iteration scatter shown in Figure 2 for the next 7 days, and obtain the specific results of each category every day from July 1st to July 7th, as shown in Table 5 and Table 6.

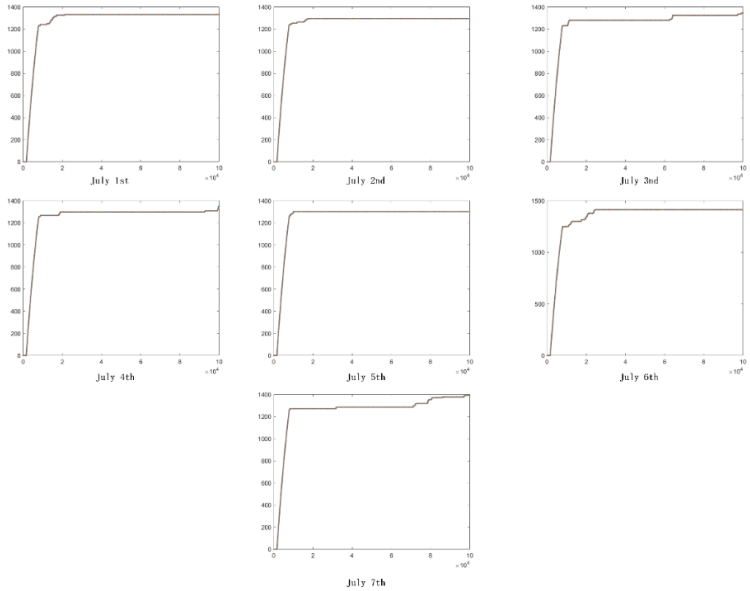

Figure 2: The total profit scatter of each category from July 1st to July 7th.

Table 5: Optimal Pricing Strategy from July 1 to 7, 2023 (¥/kg)

Category	**1st day**	**2nd day**	**3rd day**	**4th day**	**5th day**	**6th day**	**7th day**
Cauliflower	11.43	10.95	10.53	11.70	11.29	12.33	11.04
Flower leaves	8.22	8.22	8.18	8.22	8.22	8.22	8.22
Eggplants	9.47	9.34	10.11	9.47	9.01	9.33	9.36
Chilies	7.92	7.94	8.12	7.66	7.86	7.95	7.87
Edible fungi	7.25	7.57	8.76	7.42	7.44	7.54	8.34
Aquatic root vegetables	13.21	12.74	12.87	12.97	12.73	13.03	14.47

Table 6: Optimal replenishment volume from 1 to 7 July 2023 (kg)

Category	**1st day**	**2nd day**	**3rd day**	**4th day**	**5th day**	**6th day**	**7th day**
Cauliflower	34.64	36.04	37.83	34.63	35.31	33.15	35.81
Flower leaves	167.18	164.85	162.39	170.32	167.61	172.84	166.57
Eggplants	75.39	72.84	67.20	75.01	75.51	79.52	73.04
Chilies	58.85	57.77	58.05	58.32	57.76	58.61	61.54
Edible fungi	59.37	57.15	52.84	59.24	57.50	59.31	51.66
Aquatic root vegetables	18.77	19.52	19.97	19.32	19.26	19.40	14.71

The optimal pricing strategy and optimal replenishment were solved according to Table 5 and Table 6, and meanwhile, the maximum daily profit and corresponding sales from July 1 to July 7, 2023, could also be solved, as shown in Table 7.

Table 7: Maximum Profit and corresponding Sales from July 1 to 7, 2023 (¥)

	1st day	2nd day	3rd day	4th day	5th day	6th day	7th day
Profit maximum	1332.3	1295.6	1346.2	1348.4	1301.5	1413.3	1391.5
Corresponding sales	3272.4	3219.3	3244.9	3293.5	3231.5	3370.6	3225.7

7. Conclusions

The correlation between vegetables was analyzed. It was found that the correlation was mostly moderate and only the correlation between cauliflowers and flower leaves was strongly positive.

According to the final replenishment, pricing strategy, and maximum profit, the fluctuations of replenishment and pricing of each category were small, among which the replenishment of flower leaves remained the largest and the pricing of aquatic root vegetables remained the highest. The maximum profit of the supermarket in the next 7 days generally remained above 1,300. Despite a slight fluctuation in the number of days, the overall income was still considerable.

A combined research idea and framework were provided for the field of marketing, and the final results had a certain reference value for the marketing strategy of supermarkets. It enabled the supermarket to create as much profit as possible under the condition of fixed funds, avoiding the waste of commodities due to market and other causes.

References

[1] Han Xianjun, Liu Yanli, Yang Hongyu. Multiple linear regression-guided stereo matching algorithm [J]. Journal of Computer-Aided Design & Computer Graphics, 2019, 31 (1): 84-93.

[2] Cui Yibin, Liu A, Ding A, Chao A, Zheng A. Analysis of pollution sources in Chang tan Reservoir, Zhejiang Province based on absolute principal component-multiple linear regression (APCS-MLR) model[J]. Journal of Ecology and Rural Environment, 2023, 39 (4).

[3] Zhang M. Forecasting of runoff in the lower Yellow River based on the CEEMDAN-ARIMA model [J]. WATER SUPPLY, 2023, 23 (3).

[4] Yu Z. ARIMA Modelling and Forecasting of Water Level in the Middle Reach of the Yangtze River [M]. 4th International Conference on Transportation Information and Safety (ICTIS), 2017.

[5] Phillips C, Hoenigman R, Higbee B. Food Redistribution as Optimization [J]. 2011. DOI:10.1371/journal.pone.007553.

[6] Zhou Junjie. Annealing inversion of geomagnetic simulation considering prior information [J]. Geological Review, 2023, 69 (S01).

[7] Yang Haoxu. Correlation analysis of vegetable price and sales volume——Taking oil wheat as an example [J]. Food Safety Guide, 2018 (21).