Research on Automated Pricing and Supplementation Decision-making of Vegetable Products

Zejie Tian*

School of Information and Communication Engineering, Communication University of China, Beijing, China, 100024

*Corresponding author: tianzejiecuc@gmail.com

Abstract. In fresh supermarkets, the shelf life of vegetable products is usually very short, so it is crucial to replenish inventory daily based on historical sales data and product line requirements. In order to strengthen decision-making and inventory management, this article first selects representative supermarket sales data and conducts appropriate preprocessing, and then adopts a series of effective analysis methods. Firstly, using Kappa consistency testing to explore the correlation between different vegetable categories can help determine which products may be correlated in sales. In addition, in order to better formulate pricing and replenishment strategies for vegetable products, this article introduces a cost-plus pricing model. When determining the markup coefficient, multiple linear regression was used to ensure that the price-setting process is reasonable and consistent with cost and demand. In order to further confirm the accuracy and feasibility of this method, the article calculated the R-squared value of multiple linear regression, which reached 0.975, providing strong validation for the effectiveness of this method. When formulating replenishment strategies, in order to obtain sales forecasts for vegetable categories within the next 7 days, this article also uses a seasonal ARIMA model, which helps supermarkets plan inventory and supply chain more effectively. These strengthening measures have jointly improved the decision-making ability of vegetable product management in supermarkets, promoted more effective data-driven strategies, ensured the freshness and quality of vegetable products, and maximized the economic benefits of supermarkets.

Keywords: Kappa Consistency Testing, Multiple Linear Regression, Seasonal ARIMA Models, Big Data.

1. Introduction

In the management of fresh food supermarkets, vegetable products are characterized by their short shelf life and the tendency for their quality to deteriorate over time. Therefore, supermarkets need to replenish their stock promptly every day based on historical sales and demand, even though procurement transactions typically occur between 3:00 AM and 4:00 AM. Supermarket owners are required to make restocking decisions without knowing the specific individual products and their procurement prices. Typically, supermarkets use a cost-plus pricing method and offer discounts on products with transportation losses or lower quality.

Market demand analysis is crucial for restocking and pricing decisions in this context. Demand analysis can consider the relationship between sales volume and time as well as supply aspects. Furthermore, due to the pronounced seasonality in vegetable supply during different seasons, seasonal factors also need to be taken into account.

In previous work, someone used Pearson linear correlation [1] analysis or Spearman correlation [2] analysis to examine data consistency and employed LSTM [3] for single-item sales forecasting. However, based on this paper's approach, this paper first noticed that the data is not suitable for linear models, as vegetable sales data is typically discontinuous and does not follow a normal distribution. Therefore, this paper employed Kappa analysis to study the consistency among different products. Additionally, this paper used a multiple linear regression model to estimate sales prices, considering the relationship between cost markup ratios, sales volume, loss rates, and wholesale prices. For sales forecasting, this paper opted for a seasonal ARIMA model to better capture the seasonality in sales

data. The goal of these methods is to more accurately predict sales demand and formulate reasonable pricing strategies.

2. Methods

2.1. The basic fundamentals of Kappa consistency testing

Kappa consistency analysis [4], often referred to as Cohen's Kappa, is a statistical measure used to assess the level of agreement between two or more raters or observers when categorizing or coding items into discrete categories. This analysis is particularly valuable in fields such as interrater reliability assessment, medical diagnosis, and content analysis.

The primary purpose of Kappa consistency analysis is to determine whether the observed agreement between raters is higher or lower than what would be expected due to chance alone. In other words, it helps researchers and professionals evaluate whether the agreement between raters is statistically significant.

Kappa values range from -1 to 1, with specific interpretations:

A Kappa value of 1 indicates perfect agreement beyond chance.

A Kappa value of 0 suggests that the observed agreement is equivalent to what would be expected by random chance.

A Kappa value less than 0 indicates that there is less agreement than would be expected by random chances.

In research and practical applications, a Kappa value above 0.7 is often considered substantial agreement, while values below 0.4 suggest poor agreement. Kappa consistency analysis is a valuable tool for ensuring the reliability and validity of data collection processes, ultimately leading to more robust research outcomes and decision-making.

The formula for Kappa (κ) consistency test is as follows:

$$Kappa(\kappa) = \frac{p_o - p_e}{1 - p_e}, \tag{1}$$

Where: p_o(Observed Agreement) is the actual level of agreement between observers, representing the frequency of actual agreement among observers. p_e (Expected Agreement by Chance) is the agreement that would be expected by random chance, typically estimated by calculating the expected agreement among observers. p_e is often calculated based on the distribution of ratings by each observer and the overall distribution.

2.2. The basic fundamentals of Multiple Linear Regression

Multiple linear regression [5][6] is a statistical technique used to analyze the relationship between a dependent variable (Y) and two or more independent variables ($X_1, X_2, X_3, \ldots X_k$). It aims to establish a linear equation that best represents how the independent variables collectively influence the dependent variable.

These steps guide the process of conducting multiple linear regression analysis to understand and model the relationships between variables without including mathematical formulas in the instructions.

The formula for multiple linear regression is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k + \varepsilon, \tag{2}$$

Where: Y is the dependent variable. $X_1, X_2, X_3, \ldots X_k$ are the independent variables. β_0 is the intercept (constant term). β_1, β_2, β_3, ..., β_k are the coefficients of the independent variables. ε is the error term.

The goal of multiple linear regression is to determine the values of the coefficients (β_0, β_1, β_2, β_3, ..., β_k) that best fit the data and allow us to make predictions or infer the relationship between the independent and dependent variables.

2.3. The basic fundamentals of the Seasonal ARIMA Model

The Autoregressive Integrated Moving Average (ARIMA) model [7-9] is a widely used time series forecasting method. It combines autoregression (AR) [10], differencing (I for integrated), and moving averages (MA) to capture different aspects of time series data.

Here are the analysis steps of Seasonal ARIMA Model:

Decompose the time series into trend data, seasonal data, and random data to preliminarily assess the seasonality effect of the data.

Seasonal ARIMA models require that the time series pass a stationarity test. If $P < 0.05$, it indicates that the series is stationary. If the original time series does not meet stationarity, perform differencing and seasonal differencing until the series becomes stationary.

Examine the final differenced series plot and conduct autocorrelation analysis on the time series to estimate the values of p and q based on the truncation pattern.

ARIMA models require that the model exhibits pure randomness, meaning that the model residuals are white noise. Examine the model diagnostic table and assess the white noise property based on the P-value of the Q-statistic (P-value greater than 0.05 indicates white noise). Additionally, analyze the AIC and BIC values as lower AIC and BIC values indicate better model fit. Lastly, evaluate the model's goodness of fit the series by looking at the p^2value, where values closer to 1 indicate better model performance.

3. Results

3.1. Data sampling

The data for this paper is sourced from a large fresh produce retail chain located in Beijing. It covers detailed sales records for the entire year of 2022, with data available for each day. This supermarket benefits from a prime geographical location and attracts a diverse customer base from various segments of society. Consequently, it effectively represents the characteristics of the entire fresh food supermarket industry.

The selection of this particular supermarket was deliberate to ensure that the sample used in the study is highly representative and accurately reflects the sales dynamics within the fresh food supermarket sector. Sales data has been meticulously recorded and organized in an Excel spreadsheet, including information such as sales dates, prices, and product codes. Furthermore, to facilitate a comprehensive multiple linear regression analysis, this paper also collected data on the spoilage rates for various product categories. The diverse sources and nature of this data provide a robust foundation for the research, ensuring the reliability and comprehensiveness of the analysis.

3.2. Correlation Analysis

After collecting the data, the data was preprocessed. Firstly, a classification function was written using MATLAB language, and six types of vegetables were classified based on product codes for 193 vegetables in the Excel table, and then a graph was created. Figure 2 shows the normalized weekly sales volumes of six types of vegetables in 2022.

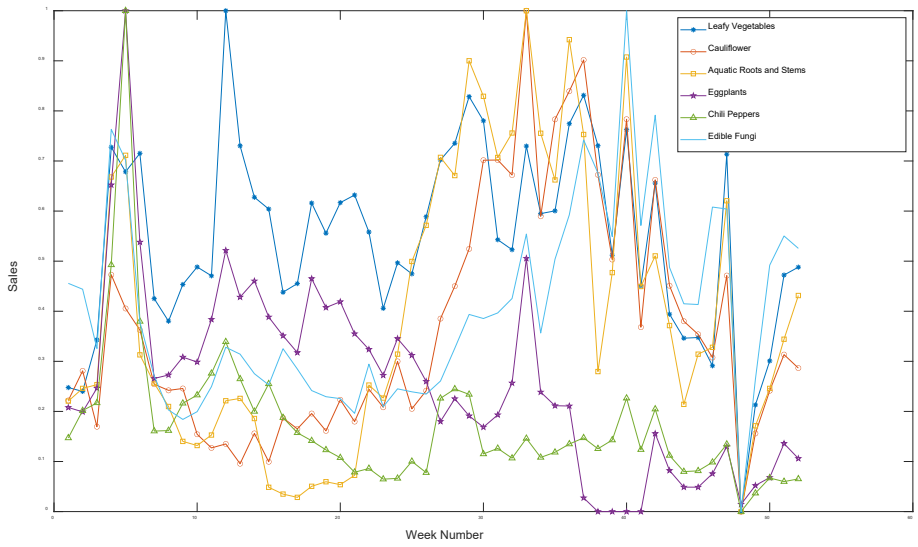

Figure 1 Weekly Total Sales of 6 Vegetable Categories (Normalized)

From Figure 1, it can be observed that there is consistency and correlation in the sales volume variations throughout the year for chili, cauliflower, and aquatic root vegetables.

Next, this paper will analyze the correlation between different product categories. Since the data does not meet the requirements of linear data, continuity, and normal distribution, it is not suitable to use Pearson linear correlation analysis or Spearman correlation analysis. This paper will test whether there is a statistically significant relationship between different variables by assessing whether the p-value shows significance ($P<0.05$). If it is significant, it indicates a correlation between the two variables. This paper will then analyze the direction and degree of correlation using the Kappa coefficient. The resulting correlation between some individual products is as follows:

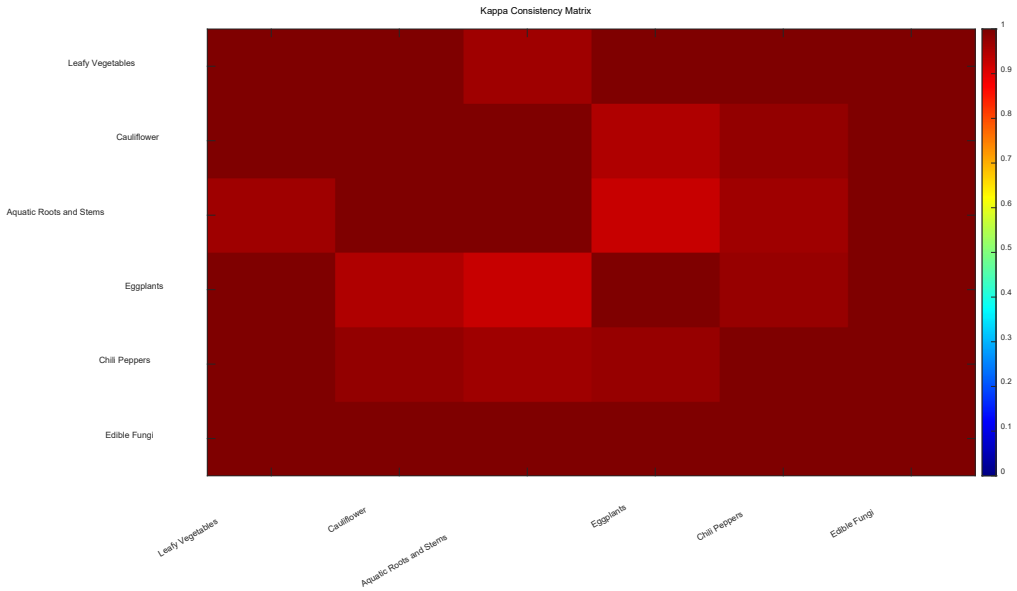

Figure 2 The Kappa Consistency Matrix of the products

Based on this paper's analysis, this paper has identified several key insights worth highlighting. As depicted in Figure 2 above, there are distinct products within this paper's dataset that display a remarkable degree of correlation. This intriguing finding raises questions about the driving factors behind these strong associations.

One hypothesis to explain these correlations is the concept of product complementarity. It appears that customers often opt to purchase certain products together because they complement each other,

either in flavor or use. For example, the co-purchase of cauliflower and cauliflower leaves suggests that customers may buy these items to create culinary dishes or recipes that utilize both components. This phenomenon not only showcases the customers' preferences but also highlights potential opportunities for marketing and bundling strategies within the fresh food supermarket industry.

Understanding the nuances of these product correlations can have a profound impact on inventory management, pricing strategies, and customer satisfaction. It provides valuable insights into customer behavior and preferences, enabling supermarkets to make data-driven decisions that enhance their offerings and overall shopping experience. Further research and analysis in this area can uncover even more opportunities for the fresh food retail sector.

3.3. The relationship between the cost markup ratio and sales volume, loss rate, and wholesale price

First, according to the cost markup pricing method, let's assume the pricing model as:

$$P = C(1 + a) \tag{3}$$

where: P represents pricing, C represents cost, and α represents the cost markup coefficient.

Considering that there is a multivariate linear relationship between the cost markup coefficient and sales volume, loss rate, and wholesale price, this paper establishes a multivariate linear regression model for the cost markup coefficient about sales volume, loss rate, and wholesale price as follows:

$$a = K_1 S + K_2 A + \varepsilon \tag{4}$$

where: ε represents other coefficient influencing factors. K_1, K_2represent normal coefficient.

After performing statistical analysis and processing with the data from section 3.1, this paper can use the example of fitting the coefficient for chili to illustrate the detailed analysis and solution process. Using the Partial Least Squares (PLS) [11] method, the steps to obtain the model are as follows:

1. Obtaining the Composition of Main Components: The component matrix is a vital tool in understanding the intricate relationships between the main components and the original variables. Each main component is essentially a linear combination of the original variables, and the component matrix reveals precisely how each original variable contributes to these main components. This matrix is indispensable in gaining insights into the underlying structure of the data and identifying which variables have the most significant influence on the main components.

2. Determining Variable Importance via Factor Loading Coefficients: The factor loading coefficient table provides valuable insights into the significance of each original variable about the main components. These loading coefficients effectively quantify the correlation between the original variables and the main components. Larger loading coefficients signify that the respective original variables wield greater influence over the main components. This aspect is crucial in pinpointing which variables play a pivotal role in explaining the variability captured by the main components.

3. In the context of Partial Least Squares Regression (PLSR), the culmination of these analyses leads to the development of a standardized formula for the regression model. This model, which incorporates the main components and their respective weights (loading coefficients), forms a multivariate regression framework.

The final fitting results are shown in Table 1:

To evaluate the performance of multiple linear regression, this paper first used MATLAB to generate curves for fitting data and real data, as shown in Figure 3. By visually observing the graphics, readers can clearly understand the quality of the fitting results. In addition, to comprehensively evaluate the performance of the model, this article also calculated the R-squared value, which is used to measure the degree to which the model interprets the variability of the dependent variable. The value range of the R-squared value is usually between 0 and 1, and the closer the value is to 1, the higher the fit of the model.

Table 1 Cost markup ratio for each category

6 Categories	Model
Chili Peppers coefficient α_1	$a_1 = -26.041 - 3.088A - 0.036S$
Eggplants coefficient α_2	$a_2 = 85.306 - 11.881A + 0.081S$
Edible Fungi coefficient α_3	$a_3 = -57.771 + 7.526A - 3.646S$
Cauliflower coefficient α_4	$a_4 = -38.183 + 2.756A - 0.611S$
Leafy Vegetables coefficient α_5	$a_5 = 76.151 - 7.297A - 1.163S$
Aquatic Stems coefficient α_6	$a_6 = -87.07 + 7.318A - 0.233S$

In the study of this paper, the calculated R-squared value is 0.975, indicating that the multiple linear regression model in this article is excellent and successfully fits the original data. This high R-squared value confirms the superiority and accuracy of the model, providing solid support for the research in this paper.

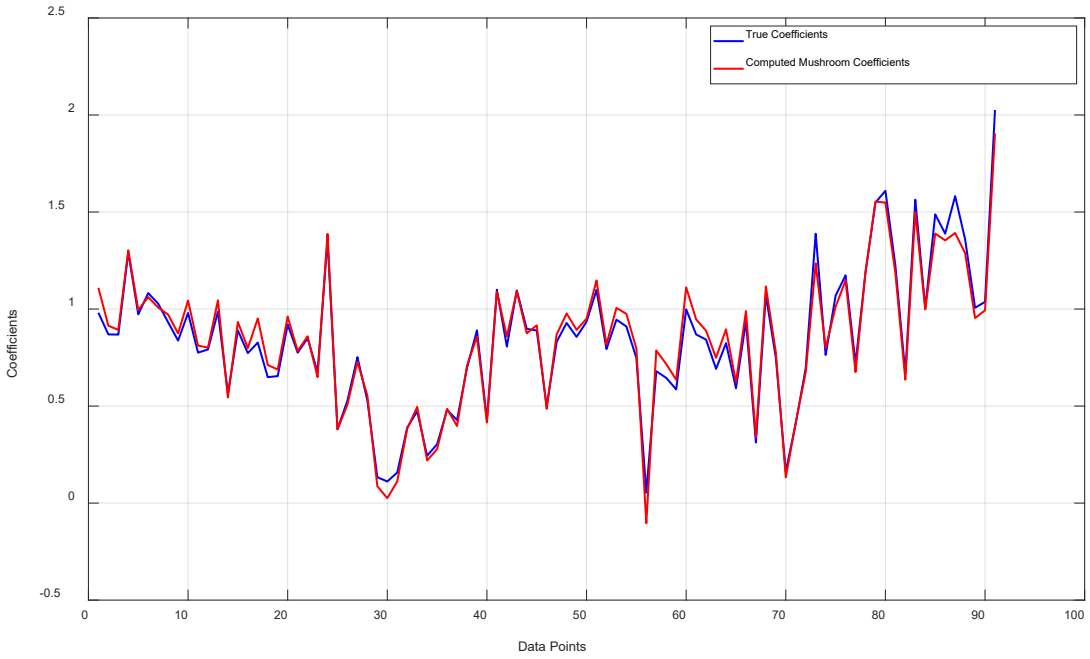

Figure 3 The Curves of fitting data and real data(take Mushroom as an example)

3.4. Sales price prediction based on seasonal ARIMA model from July 1, 2023 to July 7, 2023

Acquiring daily sales data for the six product categories spanning from April 1, 2023, to June 30, 2023, provides us with a rich dataset for analysis. A noteworthy observation is that this data exhibits a recurring pattern, manifesting as periodic changes with a 7-day cycle. This cyclic behavior can significantly impact sales patterns and warrants a specialized approach to forecasting.

In light of this, this paper has chosen to employ a Seasonal ARIMA (Auto Regressive Integrated Moving Average) model to predict sales data for the upcoming week, spanning from July 1, 2023, to July 7, 2023. The utilization of the seasonal ARIMA model takes into account the recurring nature of the data, enabling us to make informed predictions for the specified timeframe.

This paper's forecast results, which are presented in Table 2-3, are a testament to the power of statistical analysis and predictive modeling. By analyzing past data and employing advanced modeling techniques, this paper is better equipped to anticipate sales trends and make informed decisions, ultimately contributing to more effective product management and business strategies. It's important to note that similar methodologies were applied to obtain forecasts for the remaining product categories, ensuring a comprehensive and data-driven approach to decision-making.

Table 2 Prediction results of Leafy Vegetables

Date	Sales Forecast	Wholesale Forecast	Loss Rate	Cost Markup Pricing
7.1	192.02	3.34	10.28	655.94
7.2	173.51	2.96	10.28	675.22
7.3	128.54	3.27	10.28	897.24
7.4	147.36	3.72	10.28	983.13
7.5	171.09	3.71	10.28	862.05
7.6	155.09	3.49	10.28	893.23
7.7	168.02	3.58	10.28	852.23

Table 3 Prediction results of Cauliflower

Date	Sales Forecast	Wholesale Forecast	Loss Rate	Cost Markup Pricing
7.1	30.31	5.17	14.14	255.83
7.2	27.79	5.05	14.14	250.74
7.3	18.49	5.23	14.14	150.58
7.4	18.47	5.54	14.14	120.21
7.5	18.45	5.33	14.14	140.76
7.6	19.59	5.37	14.14	145.57
7.7	20.59	5.21	14.14	171.06

By closely examining the data presented in Table 2 and Table 3, it becomes evident that various product categories exhibit noteworthy distinctions across a range of parameters. These disparities are a result of the comprehensive influence of several factors, contributing to the diversity observed among the categories. These parameters encompass aspects such as sales volume, wholesale pricing, loss rates, and cost markup pricing, all of which play pivotal roles in defining the distinct characteristics of each product category. This multifaceted analysis sheds light on the intricate dynamics at play within the realm of product categories and underscores the significance of considering these variations in decision-making processes.

4. Conclusions and outlooks

In the context of fresh food supermarkets, a significant challenge arises from the short shelf life of most vegetable products and their susceptibility to quality deterioration over time. These products often cannot be carried over to the next day if left unsold. To address this issue, supermarkets have adopted a daily replenishment strategy, relying on historical sales data and specific demand patterns for each product category. In this paper's effort to optimize vegetable product management, this paper conducted an in-depth analysis of sales data. This paper's goal is to uncover intricate relationships among different vegetable categories, identify temporal sales patterns, and establish a robust cost-plus pricing model. This multifaceted approach empowers us to make informed decisions that enhance the management of vegetable products.

The model effectively addresses pricing concerns by optimizing discount strategies for revenue maximization. It thoroughly explores ordering, pricing, and discounting tactics, utilizing various time dimensions (monthly, weekly, daily) to analyze sales data and understand trends. However, there are some challenges to consider, such as potential errors introduced when using an average loss rate for product categories due to their shared characteristics. Additionally, the incorporation of complex mathematical and statistical methods in the model can result in a computationally intensive process. To further enhance the model, integrating inventory data may be explored to apply it to short-lifecycle products and temporary price discount scenarios. Advanced optimization algorithms can also be

considered to handle complex constraints and larger-scale problems, ultimately improving decision-making across various scenarios.

References

[1] Senthilnathan S. Usefulness of correlation analysis[J]. Available at SSRN 3416918, 2019.

[2] Wang Z, Lv J, Tan Y, et al. Temporospatial variations and Spearman correlation analysis of ozone concentrations to nitrogen dioxide, sulfur dioxide, particulate matter and carbon monoxide in ambient air, China[J]. Atmospheric Pollution Research, 2019, 10(4): 1203-1210.

[3] Yu Y, Si X, Hu C, et al. A review of recurrent neural networks: LSTM cells and network architectures[J]. Neural Computation, 2019, 31(7): 1235-1270.

[4] Foody G M. Explaining the unsuitability of the kappa coefficient in the assessment and comparison of the accuracy of thematic maps obtained by image classification[J]. Remote sensing of environment, 2020, 239: 111630.

[5] Rath S, Tripathy A, Tripathy A R. Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model[J]. Diabetes & metabolic syndrome: clinical research & reviews, 2020, 14(5): 1467-1474.

[6] Ciulla G, D'Amico A. Building energy performance forecasting: A multiple linear regression approach[J]. Applied Energy, 2019, 253: 113500.

[7] Liu X, Lin Z, Feng Z. Short-term offshore wind speed forecast by seasonal ARIMA-A comparison against GRU and LSTM[J]. Energy, 2021, 227: 120492.

[8] Dimri T, Ahmad S, Sharif M. Time series analysis of climate variables using seasonal ARIMA approach[J]. Journal of Earth System Science, 2020, 129: 1-16.

[9] Noureen S, Atique S, Roy V, et al. Analysis and application of seasonal ARIMA model in Energy Demand Forecasting: A case study of small scale agricultural load[C]//2019 IEEE 62nd International Midwest Symposium on Circuits and Systems (MWSCAS). IEEE, 2019: 521-524.

[10] Lenza M, Primiceri G E. How to estimate a vector autoregression after March 2020[J]. Journal of Applied Econometrics, 2022, 37(4): 688-699.

[11] Purwanto A. Partial least squares structural squation modeling (PLS-SEM) analysis for social and management research: a literature review[J]. Journal of Industrial Engineering & Management Research, 2021.