Information Systems and Economics (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/infse.2024.050120
ISSN 2523-6407 Vol. 5 Num. 1

Research on Pricing and Replenishment of Vegetable Products Based on the Optimization of Supermarket Revenue

Keding Sheng, Xingshuai Mei[#], Xitong Yang[#], Huimin Zhang[*,#]

School of Politics and Public Administration, Qinghai Minzu University, Xining, 810007, China
[*]Corresponding author: Huimin_0608@163.com
[#]These authors contributed equally.

Keywords: Association Rule Learning, Pearson Correlation Coefficient, LSTM Time Series Analysis, Knapsack Model

Abstract: This study aims to enhance the operational efficiency and sales revenue for fresh food supermarkets through preprocessing and in-depth analysis of vegetable sales data. The research begins by clarifying the relationships between individual items and product categorizations, utilizing data mining techniques such as association rule learning to reveal purchasing patterns among different items. Subsequently, the distribution patterns of individual item sales and their correlation with category sales are analyzed, employing descriptive statistics and visualization methods, such as box plots and bar charts, to test for normality within the data. Based on these findings, either the Pearson or Spearman correlation models are selected for quantitative analysis. Through data analysis, a linear regression model is constructed to delineate the relationship between total sales volume and cost-plus pricing. This model is integrated with LSTM time series analysis to guide the formulation of replenishment plans and the optimization of pricing strategies. Additionally, the relationship between product loss rates and shelf life is considered. A knapsack model is applied to optimize replenishment decisions, ensuring the maximization of individual item profit margins within the constraints of knapsack capacity, thereby improving the accuracy and economic benefits of restocking. The methodologies and models presented in this research hold significant theoretical and practical implications for the management and operation of fresh food supermarkets.

1. Introduction

In the supermarket retail industry, especially for perishables like vegetables, implementing intelligent pricing and inventory management strategies is of utmost strategic value. Unlike general products, the most notable characteristic of perishable goods is their ease of deterioration and consumption, and the value of vegetable products will undoubtedly decline over time. [1] The supermarket's management strategy for vegetables must fully consider the rate of quality degradation, the impact of seasonal and cyclical factors, and unique trading cycles. Production uncertainty and cost uncertainty are common issues in supply uncertainty, where the former has a significant impact

on the manufacturer's production and the matching of product supply with demand, while the uncertainty in production costs affects the manufacturer's pricing decisions and the retailer's ordering strategy [2]. In terms of pricing strategies, cost-plus pricing needs to be sensitively adjusted to offset the value loss brought about by quality decline. To maintain profitability and market competitiveness, retailers must find a balance between holding a reasonable inventory level and timely price adjustments. Particularly for a company producing perishable goods, the inventory management of perishable vegetable goods becomes of greater importance [3]. Overstocking can lead to spoilage and capital occupation, while understocking risks losing sales opportunities and affects customer satisfaction. By describing the spoilage characteristics of fresh agricultural products and considering costs such as expiry, damage, out of stock, ordering, and holding, a fresh agricultural product inventory cost control model is established from the supply chain perspective [4]. Considering the continued impact of the decline in vegetable quality on sales and customer satisfaction, and considering the supermarket's loss aversion, an expected utility maximization model for the supermarket has been constructed [5].Supermarket managers must use strategies based on quantitative algorithms and predictive models for decision-making. Tools like option pricing, inventory theory, and time series analysis can help predict and optimize price and inventory management. (The data source for this article: mcm.edu.cn)

2. Research on the Distribution Patterns and Interrelationships between Vegetable Categories and Individual Item Sales

During the sales analysis of vegetable products, a thorough investigation into the potential correlations and distribution patterns of sales among different categories or individual items is critical. Such analyses aid in understanding the complex dynamics of the vegetable market and consumer purchasing behavior. To achieve this goal, it is first necessary to collect and extract sales data from various vegetable categories as the basis for research. Descriptive statistical analysis of these data is implemented to mine the fundamental sales characteristics of vegetable categories and provide intuitive visualization through means such as box plots. In addition, the normality of the data needs to be assessed through appropriate testing methods, an important step determining the statistical approach to be used in subsequent correlation analyses. When data distributions exhibit normality, the Pearson correlation coefficient is used to measure the linear relationships between variables; conversely, if data distributions do not adhere to normality, the Spearman's rank correlation coefficient is utilized to assess the monotonic relationships between variables. Ultimately, this yields insights into the distribution patterns and interrelationships of sales among vegetable categories.

2.1 Descriptive Statistics and Sales Distribution Analysis

Dataset evaluation and preprocessing prior to statistical analysis are key steps in the research, ensuring the accuracy and reliability of the analysis. In this study, individual vegetable items are classified into six major categories for a more systematic understanding and analysis of market data. To ensure the quality of the statistical analysis, box plots are employed first to identify and eliminate outliers in the data, retaining data points within a reasonable range. As shown in Fig 1, only a portion of records in the dataset present outliers. By removing these outlier data, subsequent analyses are safeguarded against potential misleading influences. After outlier removal, further observation reveals that there are varying counts of missing values among different features. In such instances, appropriate imputation of these missing values is deemed necessary to prevent biased or distorted analysis results. After cleansing and processing, statistics for the six main categories of vegetable products show the following market share ratios: Florifolias 42.2%, Chili peppers 19.4%, edible fungi 16.2%, florescent vegetables 8.9%, Aquatic rhizomes 8.6%, and Solanaceae 4.8%.

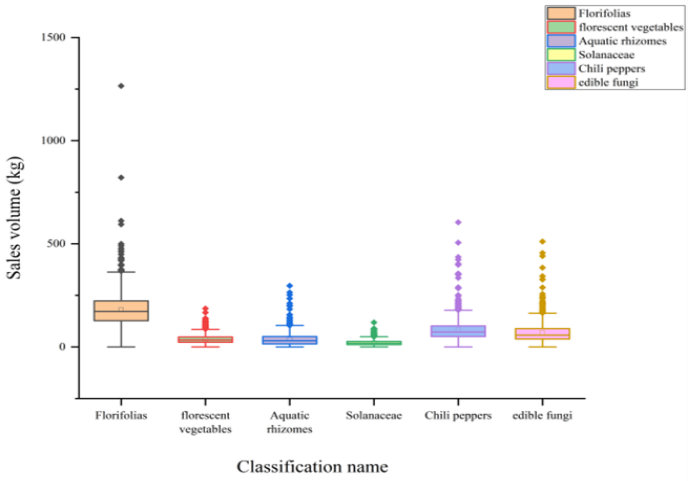

Figure 1: Box plot of sales unit price

In the categorization of data types, this paper treats the names of individual items and category designations as qualitative or categorical variables, while total sales are regarded as quantitative variables, quantifying their numerical fluctuations. To overcome the issue of disparate units among various variables, the study implemented a normalization process for the total sales, aiming to eliminate the impact of differing dimensions on statistical outcomes and to ensure the normalization of data, laying a foundation for multivariate analysis.

The normality test is a critical step in this research, as many statistical analysis methods, particularly parametric ones, operate on the presumption that the data follows a normal distribution. The study conducted a comprehensive evaluation of the data's normality by utilizing significance P-values, quantitative assessments of kurtosis and skewness, as well as graphical methods. Herein, the P-value—indicating the probability of the observed outcome—was set at a threshold of 0.05 to determine whether the data distribution satisfies the assumption of normality. For those datasets that do not follow a normal distribution, the research considered employing suitable data transformation techniques, such as logarithmic transformations, to make the data distributions more aligned with normality. For instance, in the normality test of wholesale price data, the significance value from the Kolmogorov-Smirnov (K-S) test was 0.07. Since this value is above the generally established threshold of 0.05 for significance tests, there is insufficient evidence to reject the null hypothesis, implying that there is not adequate basis to assert that the dataset does not adhere to a normal distribution. Thus, the null hypothesis is accepted, which is to say that the wholesale price data fits a normal distribution with a mean of 3.06 and a standard deviation of 11.97.

2.2 Association Analysis

While investigating the data across six major categories of vegetable products, the scatterplot matrix provided an intuitive multivariate exploration tool for this paper. Through such graphic representation, one can clearly observe the relationships between different product classifications, identify potential patterns of correlation, and pinpoint the characteristics and potential outliers of data distributions. As evidenced by Fig 2, it is noted that some vegetable product categories exhibit a degree of linear relationship, indicating that the numeric changes in one category tend to correspond proportionately to changes in another. The identification of such linear relationships provides a basis for understanding the interactions among different categories of vegetable products. Further observation revealed that the data for some types of vegetables displayed characteristics of an approximate normal distribution along the distribution of diagonal elements in the scatterplot matrix.

The form of a normal distribution suggests that the market supply and demand for these commodities may follow certain natural laws or stable economic mechanisms.

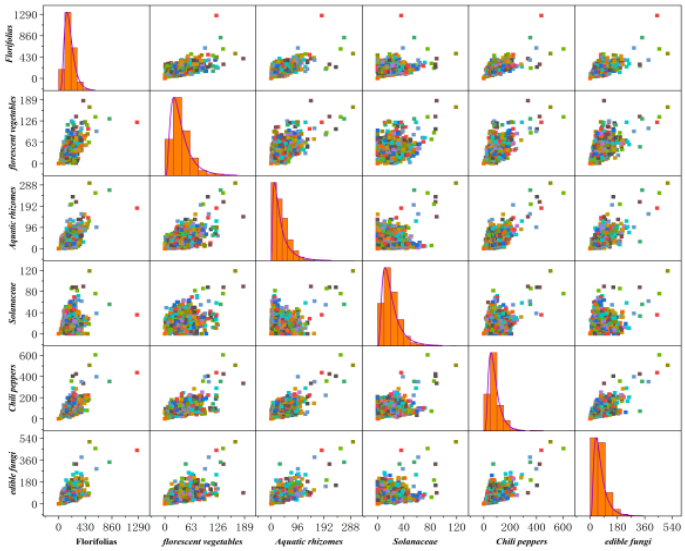

Figure 2: Scatter plot of six categories of product matrices

In the analysis of retail data sets, the application of association rule mining techniques can provide an in-depth understanding of commodity sales patterns. In the combinatory analysis directed at vegetable product categories, a key discovery was the identification of significant associations between specific vegetable products. For instance, the study revealed strong associations between Solanaceae and Chili peppers as well as between edible fungi and Chili peppers. Within this study, mining results indicated a lift of 1.333, suggesting that if a customer purchases one vegetable product, the likelihood of them choosing to buy a related vegetable product is 1.333 times the probability of making a random selection. This infers that the positive correlation implies that the sales of specific vegetable commodities are not independent of each other but rather mutually influential, a connection likely driven by consumers' purchasing habits and preferences.

2.3 Distribution Patterns of Sales Volume for Various Vegetable Categories and the Associations between Vegetable Categories

While analysing the total sales volume of individual items within the retail data set, this article applied normality tests to assess the shape of the data distribution. The histogram results of the normality tests suggest that while the data distribution is not a perfect normality, it exhibits the fundamental characteristics of a bell-shaped curve, indicating that the data distribution generally conforms to a normal distribution. Additionally, the analysis of the Q-Q plot revealed that the data points for individual item total sales volume demonstrate a high degree of fit with the reference line, further corroborating the normal distribution attributes of the data.

Analysis of sales volume trends over time allows for a better understanding of seasonal consumption patterns among different categories, and the formulation of corresponding sales strategies and promotional campaigns based on these trends. Having ascertained the normal distribution nature of the data, the paper employed the Pearson correlation coefficient model to quantitatively evaluate the strength of associations between different indicators. For a thorough analysis, the article utilized SPSS statistical software to conduct correlation tests and created a heatmap of the six major categories of vegetable products as displayed in Fig 3demonstrating their interactions.

Two types of association patterns may be observed from Fig 3. The first is a high positive correlation: there is an extremely high degree of correlation between the cauliflower types and leafy types. Similarly, a strong correlation exists between the edible fungi category and the aquatic rhizomes category. This implies that an increase in sales volume for "edible fungi" may likely lead to an increase in sales volume for "aquatic rhizomes" and vice versa. These results suggest that these categories are commonly consumed vegetables in people's daily diets. On the other hand, the correlation between the eggplant category and most other categories is low or non-existent. This indicates that their sales patterns may be independent, rather than being common household dishes. This could be related to the specific uses of eggplants in certain recipes or dishes.

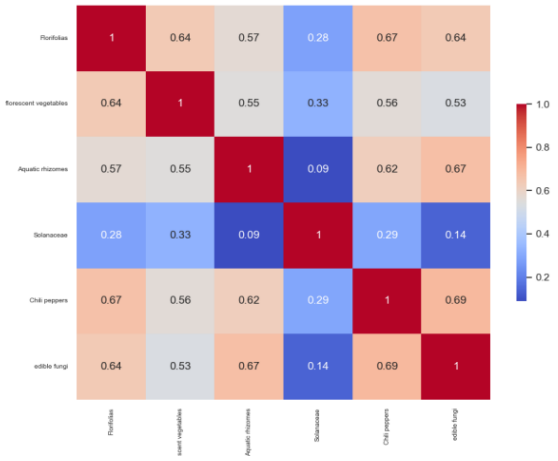

Figure 3: Heat map of six major categories of vegetable products

Analysis of the data on individual item sales and total sales leads to an understanding of the differences between various product categories. Furthermore, evaluation of the temporal trends in sales volumes for six major categories of vegetable goods reveals significant augmentations or declines in sales within specific time intervals for different categories. For instance, the ‘leafy flower’ category experiences peak sales volumes during certain periods each year, which likely correlates with particular festivals or seasonal occurrences. Some vegetables may gain popularity among consumers during certain seasons, prompting a spike in sales volume. On the other hand, for some edible fungi such as the black-skinned chicken fir mushroom, sales are observed to increase during the summer, potentially because it is the harvest season for fungi and consumers tend to purchase and consume fresh ingredients during this time.

To optimize supermarket profits through pricing and restocking strategies, the study analyzes the relationship between total sales volume and cost-plus pricing for varied vegetable categories and establishes daily restocking volumes and pricing strategies for an upcoming week (July 1-7, 2023) to maximize the profits for supermarket operators. The research summarizes sales registers and processes wholesale price data to uncover sales volumes for each vegetable category. A linear regression equation is developed to describe the relationship between sales volume and cost-plus pricing through regression analysis, and a time series forecasting method is utilized to predict sales volume for each category. The supermarket aims to define a restocking plan for single items, maintaining the inventory between 27-33 saleable items and ensuring a minimum display quantity of 2.5 kilograms per item. The study, hence, reframes the issue as a knapsack problem and, upon preprocessing, selects products predicated on profit to maximize economic returns.

3. Linear Regression

This research simplifies the model to avoid overfitting and employs the least squares method to fit

the linear relationship between total sales volume and cost-plus pricing. A non-zero overall regression coefficient for the linear regression model indicates a significant relationship among the variables. Result analysis of the F-test reveals a P-value of 0.035**, signifying statistical significance [6], allowing rejection of the null hypothesis that the regression coefficients are zero, thus satisfying the model requirements. In terms of multicollinearity, the study calculates the Variance Inflation Factor (VIF) values and identifies that all variables exhibit VIFs less than 10, suggesting no multicollinearity issues [7], indicating robust model construction.

4. Price Prediction Based on LSTM Time Series Analysis Model

Traditional econometric models, grey models, and BP neural network models—as well as fuzzy logic—have been widely applied to forecast vegetable sales volumes but possess certain deficiencies, particularly in addressing nonlinearity and long-term time series [8]. Therefore, this study adopts the LSTM (Long Short-Term Memory) time series analysis model. Assuming that daily restocking volume equates to daily sales volume, the research elects the crown daisy (Garland chrysanthemum) as the subject for analysis, employing an LSTM network for time series analysis and establishing correlations between restocking and sales volumes through trend mapping. When simulating training using MATLAB, the dataset's initial 85% serves as the training set, and the remaining 15% constitutes the test set, with iterations set at 400 cycles for training. The findings suggest crown daisies exhibit pronounced fluctuations with periodicity during restocking, tentatively identifying it as a seasonal vegetable with extensive sales volume and elongated sales cycles within seasons. To confirm the accuracy of this assessment, the study conducts validation on both the training and test sets.

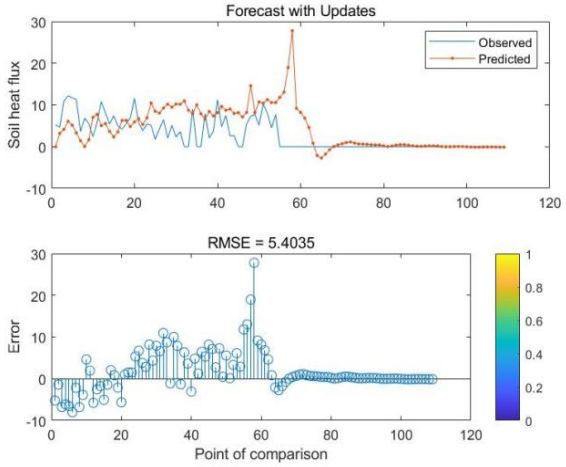

Figure 4: Comparison chart between training and testing sets

According to the results presented in Fig 4, there is a high degree of fit between the total restocking volume and the total sales volume, indicating that the LSTM time series model exhibits commendable predictive performance. Consequently, this paper employs the model to forecast the future sales volume of Garland chrysanthemum, yielding the following projected sales quantities for the upcoming week: July 1st: 3.30; July 2nd: 2.82; July 3rd: 3.38; July 4th: 3.40; July 5th: 3.62; July 6th: 4.17; July 7th: 4.45. This prediction of the Garland chrysanthemum restocking volume for the forthcoming week affirms the stability and reliability of the model. Moreover, forecasting analysis for the next week on the six major vegetable categories can yield sales results, as illustrated on the left side of Fig 5.

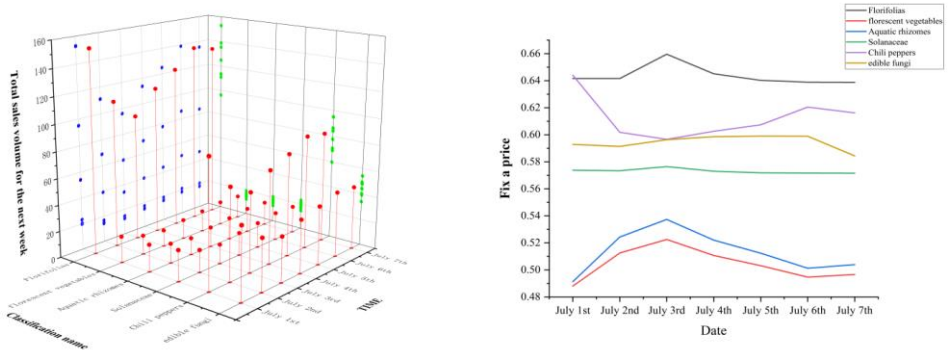

Figure 5: Sales volume and pricing strategy for the next week

By applying a regression equation to the past 30 days' data and fitting it with an LSTM, a relationship model between cost, selling price, and profit has been successfully established. The functional form linear regression model may suffer from specification errors in practical applications, and if the model's assumptions are not held, the estimations based on these assumptions could be ineffectual or even lead to erroneous conclusions [9]. Therefore, to overcome potential model errors, a penalty function was constructed using the regression equation to calculate the pricing strategy for the upcoming week. Within the right-side image of Fig 5, which displays the pricing strategy for the subsequent week, it is observed that the pricing strategy for the commodities remains relatively stable, with no significant changes. This stability aligns with everyday life characteristics and common sense, and from another perspective, this visualized data validates the accuracy and reliability of the model.

5. Knapsack Model

To maximize the profits of supermarkets, revenue calculations were conducted for 251 items, which revealed a polarization of profits with most items having a very small turnover contribution per day, earning less than 10. However, some items yield exceedingly high daily profits. Further analysis of these 251 items showed that commodities with higher sales volumes often have lower average prices, indicating an inverse relationship between average price and sales volume in the vegetable market. The restocking quantities and pricing decisions were transformed into a knapsack problem to find the optimum combination of restocking quantity and pricing to maximize profits.

To tackle the knapsack problem posed by restocking volumes and pricing decisions, the issue was transformed into a knapsack problem. Employing knapsack model theory, a greedy operator was designed to optimize and amend the initial population by considering subjective demands before objective constraints. Following that, a local search operator was developed, which improved the selection method for perturbation points to disturb and escape local optima for better-quality solutions [10]. A dynamic programming algorithm was then used to solve the problem with reduced time complexity. The analysis also filtered out data, maintaining only purchase quantities greater than 2.5kg. Additionally, a correlation was discovered between the spoilage rate of commodities and their shelf life, indicating that items with longer shelf lives tend to have lower spoilage rates. The average spoilage rate calculated from the data was found to be 9.4266%. The knapsack model calculation identified 49 selectable items, with a final selection of 36 items to be entered into the knapsack after screening for saleable items from June 24-30 and resulting in 41 items. Each category of items was calculated and placed into six knapsacks, ensuring a total of (27-33) items within each knapsack, culminating in a total profit of 755.0872585RMB.

6. Conclusion

This study has made significant progress in addressing perishable vegetable pricing and inventory management in supermarkets. The research categorized and analyzed the distribution and correlations of vegetable sales volumes, and Pearson correlation tests via SPSS software revealed the connections between vegetable categories, providing important data support for merchants' pricing and inventory management. Further, the study integrated linear regression with LSTM modeling to predict the sales trend of vegetables over the following week and introduced the knapsack problem model to optimize restocking strategies, ensuring a balance between sales volumes and costs. The research outcomes suggest that utilizing descriptive statistical analysis, association rule mining, linear regression, and time series analysis can effectively enhance the management efficiency and economic returns of perishable vegetables.

The methods and models derived in this research hold significant practical value and can assist retailers in optimizing inventory management and promotional activity effectiveness. Employing strong relationships between products for bundling or joint promotions could foster increased sales volumes and customer satisfaction. By translating restocking and pricing strategies into knapsack problem solutions, the study confirms that utilising greedy algorithms can achieve profit maximization, serving as a successful example for supermarket retailers in pricing and inventory management.

References

[1] Zhang Yajun. Comprehensive Optimization of Inventory, Price, and Quality of Perishable Products [D]. Southeast University, 2021.

[2] Lu Fen. Research on Production and Pricing Joint Decision under Supply Uncertainty [D]. Huazhong University of Science and Technology, 2018.

[3] Luo Zican, Ling Shanni. Review of foreign perishable inventory models and inventory management research [J]. Logistics Technology, 2023, 46 (07): 149-152+176.

[4] Jiao Jiao, He Lili, Zheng Junhong. DDQN-based Fresh Agricultural Products Retailer's Inventory Cost Control Model [J]. Intelligent Computer and Applications, 2023, 13 (10): 60-64+72.

[5] Pan Xiaofei, Xie Zhiheng, Wang Shuyun. Optimization Decision of Preservation Effort and Pricing of Fresh Products Supermarket Considering Loss Aversion [J]. Highway Traffic Science and Technology, 2022, 39 (06): 177-185+190.

[6] Zhi E, Wang Xueqiang, Zhang Fengfei, Xu Jingde, Ma Jun. Study on the construction of coal mine ventilation safety production standardization management system based on game combination weighting and rank sum ratio method. Journal of North China Institute of Science and Technology, 2023, 20 (03): 7-13.

[7] Xu Wenhan. Textual Information Disclosure in Annual Reports of Listed Companies: Influencing Factors and Economic Consequences [D]. Zhejiang Gongshang University, 2019.

[8] Han Jinlei, Xiong Pingping, Sun Jihong. Research on Stock Price Time Series Prediction Based on LSTM and Grey Model [J/OL]. Journal of Nanjing University of Information Science & Technology (Natural Science Edition), 1-22 [2023-11-28]

[9] Cao Ruya. Penalized Estimation of Functional Partial Linear Models under Complex Data [D]. Changchun University of Technology, 2023.

[10] Tao Lang. Study on the Optimization Method of Complex Knapsack Problem Model Based on Genetic Algorithm [D]. Anqing Normal University, 2021.