Vegetable Sales and Pricing Correlations: Insights from a RANSAC Regression Analysis

Fengyuan Yan^{1,a}, Zixian Guo^{1,b}, Yichao Yu^{1,c}, Chun Zhang^{1,d}, Xiaoyan Liu^{1,*}

[1]Shandong University of Science and Technology, Jinan, 250031, China
[a]1927614837@qq.com, [b]3127575091@qq.com, [c]1474317146@qq.com, [d]848279263@qq.com
[*]Corresponding author: lxy22602@163.com

Abstract: *The shelf life of general vegetable products in daily life is relatively short. As the sales time increases, the quality of vegetable products deteriorates. Most varieties cannot even be sold overnight. How to replenish and price the goods reasonably and timely is difficult. Extremely important and difficult. This paper develops a reasonable replenishment and pricing strategy by establishing a mathematical model and using planning methods. Using the collected sales volume and cost-plus pricing of vegetable commodities, RANSAC regression was performed to obtain the regression equation of the total sales volume of each vegetable category on the cost-plus pricing. Finally, the law of total sales volume and cost-plus pricing of each vegetable category was analyzed, and a set of equations with insignificant negative correlations was summarized.*

Keywords: *Vegetable Sales, Pricing Correlation, Regression Analysis, sales volume, cost-plus pricing*

1. Introduction

In the rapidly evolving world of fresh supermarkets, maintaining the freshness and aesthetic appeal of perishables, especially vegetables, is a formidable challenge. Vegetables, inherently, have a short shelf life. As hours elapse post-harvest, their visual appeal, a significant factor in consumers' purchase decisions, noticeably diminishes. Alarmingly, many varieties, if unsold by day's end, lose their salability by the next. This high perishability necessitates supermarkets to adopt dynamic replenishment strategies, often anchored to the historical sales and demand metrics of each product.

However, this seemingly straightforward task is mired in complexities. The vast array of vegetable varieties, coupled with their disparate origins, introduces unpredictability into the supply chain. Adding to this intricacy is the unconventional procurement time, which typically occurs in the wee hours between 3:00 and 4:00 a.m. Retailers, at this juncture, grapple with uncertainties, having to make replenishment decisions without concrete insights into product specifics and procurement prices. The prevalent pricing model for these vegetables is the "cost-plus pricing" strategy, wherein a standard markup is applied over the cost price. In scenarios where products have been compromised, either due to transport mishaps or visual deterioration, discounts are implemented to accelerate sales.

The pertinence of an astute market demand analysis cannot be overstressed. Such an analysis is pivotal for judicious replenishment and pricing determinations. Delving deeper into consumer behavior reveals that the sales volume of vegetable products often mirrors temporal patterns, hinting at discernible demand trends. Simultaneously, from a supply perspective, the period from April to October witnesses a surge in vegetable variety availability. This bounty, juxtaposed against the spatial constraints of supermarket shelves, underscores the essence of a meticulously crafted sales mix, ensuring both variety and turnover. In light of these challenges, this study aims to delve into optimal replenishment and pricing strategies for fresh supermarkets, providing actionable insights for industry stakeholders.

2. The intricacies of managing perishables

2.1 Shelf Life and Consumer Behavior

Studies by [1] emphasized the direct correlation between the visual appeal of vegetables and their shelf life. They highlighted that consumer purchase decisions are significantly swayed by the freshness and aesthetic appeal of products. This aligns with [2], where the authors discussed the psychology behind consumers' aversion to visually deteriorated vegetables, irrespective of their actual freshness.

2.2 Replenishment Strategies

The challenges associated with dynamic replenishment, especially for highly perishable items, were extensively studied by [3]. Their research underscored the necessity to base replenishment strategies on historical sales and demand metrics, accounting for variables like seasonality and harvest times.

2.3 Supply Chain Unpredictability

The unpredictability introduced by the vast array of vegetable varieties and their disparate origins was highlighted by [4]. They elaborated on how these elements, combined with unconventional procurement times, exacerbate the complexities faced by retailers.

2.4 Pricing Models

The "cost-plus pricing" strategy prevalent in fresh supermarkets was dissected by [5]. Her analysis shed light on how the pricing strategy, while being straightforward, needs to be adaptive, especially when vegetable products get compromised during transport or due to other factors.

2.5 Demand Analysis

The significance of market demand analysis in determining replenishment and pricing decisions was addressed by [6]. Their work resonated with the observation that vegetable sales volumes often reflect discernible demand trends anchored to temporal patterns.

2.6 Variety and Turnover Challenges

The balancing act between variety and turnover, especially during peak vegetable availability months, was studied by [7]. Their insights brought to the fore the challenges posed by spatial constraints of supermarket shelves.

This study builds upon these foundational works, aiming to provide actionable and comprehensive strategies for fresh supermarkets in the context of optimal replenishment and pricing.

3. The Pearson correlation coefficient

A critical component of the research is the correlation analysis between continuous variables. At this juncture, it's pivotal to underline the role of the Pearson correlation coefficient, also known as the product-moment correlation coefficient, in such analytical endeavors. Derived from a comprehensive dataset detailing the sales transactions and wholesale prices of various products from the supermarket for the period from July 1, 2020, to June 30, 2023, our correlation study leans heavily on this coefficient. As stipulated in previous research[1], the Pearson correlation is best suited for continuous data that both adheres to a normal distribution and presents a linear relationship[2]. In light of this, a preceding step of our analysis necessitates a rigorous normality test on the dataset, ensuring its alignment with the prerequisites for an effective Pearson correlation examination.

3.1 Normality test

Expanding on the context of our investigation, the validity of the analytical methods rests upon the foundational assumption of data normality. The integral part of this assessment revolves around discerning whether the collected data adheres to a normal distribution, a presumption that has significant implications for the subsequent statistical analyses.

For a more structured approach, let's delineate our hypotheses as follows:

- Null HypothesisH_0: The data follows a normal distribution.

- Alternative Hypothesis H_1: The data does not follow a normal distribution.

Considering the extensive datasets for the categories - Chili peppers, Cauliflowers, Eggplants, Leafy flowers, Edible fungi, and Aquatic rhizomes, each boasting sample sizes surpassing 5000, it's pertinent to employ the S-W test (Shapiro-Wilk test) for these variables. This rigorous examination intends to authenticate or refute our initial presumption of normality. The detailed results of this test are collated and presented in Table 1 for comprehensive clarity.

Table 1: Results of the S-W Test

Variable name	S-W test
Peppers	0.943
Cauliflower	0.954
Solanula	0.979
Flowers & leaves	0.983
Edible fungi	0.944
Aquatic rhizomes	0.948

From the above table, it can be observed that all six categories exhibit significance levels satisfying ($P > 0.05$), which indicates the lack of any significant deviation from normality. Thus, we can infer that the data adheres to a normal distribution. Furthermore, based on the visualization provided in the previous section, the differences between the experimental data are minimal. Therefore, it's appropriate to proceed with the Pearson correlation analysis methodology.

From the above table, it can be observed that all six categories exhibit significance levels satisfying { $P > 0.05$ }, which indicates the lack of any significant deviation from normality. Thus, we can infer that the data adheres to a normal distribution. Furthermore, based on the visualization provided in the previous section, the differences between the experimental data are minimal. Therefore, it's appropriate to proceed with the Pearson correlation analysis methodology.

3.2 Based on the RANSAC regression model, the relationship between total sales volume and cost-plus pricing is solved

The goal is to establish a model that describes the relationship between the total sales volume of various vegetable categories and their cost-plus pricing strategy. Once the model is established, the next step is to analyze this relationship.

External factors such as weather conditions and natural disasters can significantly affect vegetable sales. Consequently, the data samples might contain outliers or anomalies that could skew the results of our regression model, reducing its accuracy and predictive power. It is essential to address these outliers to derive meaningful insights.

The RANSAC (RANdom SAmple Consensus) algorithm offers a robust solution to this problem. Unlike conventional iterative methods, which attempt to remove outliers by repetitively refining the regression model, RANSAC adopts a unique approach. The method randomly selects subsets of the original data and constructs a series of models. The optimal model—determined by the one that aligns best with the majority of the data—is then chosen, and outliers are identified based on their deviation from this model. This approach ensures a higher degree of accuracy and resilience against outliers, improving both regression and prediction results[3].

RANSAC (RANdom SAmple Consensus) algorithm[4] is characterized by two main features: randomness and assumption. The assumption refers to the belief that the data randomly drawn in each iteration is correct. Meanwhile, the randomness signifies that the subsets of data selected during each iteration are chosen at random.

General Steps of RANSAC Regression

Step 1: Determine the required number of iterations K based on the desired proportion of inliers { θ } and a set confidence probability P.

Step 2: Use random sampling to fit the model parameters. Once the parameters are obtained, test them with all data points to determine the number of inliers n' for the proposed model.

Step 3: The optimal model is identified as the one with the highest number of inliers or, when the number of inliers is sufficiently large, the one with the minimum variance of errors.

Step 4: Once the best model is identified, re-fit it using all of its inliers to determine the final model parameters.

To further elucidate, the uniqueness of RANSAC lies in its ability to handle a significant number of outliers. By continually drawing random subsets of data and testing these subsets against the entire dataset, RANSAC identifies and discards outliers, refining the model in the process. This approach provides a higher degree of resistance against erroneous data, ensuring that the final regression model is robust and accurate.

Using SPSSPRO for linear fitting, we obtained the confidence probabilities P for the six categories. The results are displayed in Table 2 below.

Table 2: Significance P Value Table

Category	Significance	Category	Significance P Value
Flowering Cabbage	0.001***	Eggplant	0.003***
Flower and Leaf	0.000***	Edible Mushrooms	0.000***
Chili	0.043*	Edible Mushrooms	0.000***
Note: The symbols ***,**, and * correspond to significance levels of 1%, 5%, and 10%, respectively			

The values in Table 2 indicate that each category has a significant relationship with the dependent variable at their respective significance levels. Specifically, Flower and Leaf, Edible Mushrooms, and Aquatic Roots show a very high level of significance (1% level). Meanwhile, Flowering Cabbage and Eggplant also present a strong significance, and Chili is significant at the 10% level. This table provides valuable insights into the categories' relationships with the dependent variable and can guide further analyses and decisions.

According to the regression model, the linear regression equation of total sales volume and cost-plus pricing of 6 categories can be obtained.

$$\text{Sales} = \text{Constant} + \text{Cost Markup} \times \text{Cost Markup Pricing}$$

The equation represents a model predicting sales. In this, "Sales" is the total number of units sold, "Constant" signifies a baseline sales value, "Cost Markup" indicates the proportion by which the base cost of a product is increased, and "Cost Markup Pricing" is the resultant price after applying the markup to the base cost.

$$Y_{\text{cauliflower category}} = 0.553 - 0.018X_{\text{cauliflower category}} \tag{1}$$

$$Y_{\text{Mosaic leaves}} = 1.05 - 0.072X_{\text{Mosaic leaves}} \tag{2}$$

$$Y_{\text{chili}} = 0.568 - 0.012X_{\text{chili}} \tag{3}$$

$$Y_{\text{eggplant}} = 0.556 - 0.008X_{\text{eggplant}} \tag{4}$$

$$Y_{\text{Edible fungi}} = 0.568 - 0.012X_{\text{Edible fungi}} \tag{5}$$

$$Y_{\text{aquatic rhizome}} = 0.568 - 0.012X_{\text{aquatic rhizome}} \tag{6}$$

According to the mathematical model introduced above, the cost-plus pricing of the six major vegetable categories does not have a significant impact on their total sales. In order to deeply understand the various factors that affect cost-plus pricing, this study focuses on analyzing the correlation between cost and cost-plus pricing. cost-plus pricing can be defined as the sum of the total unit cost of a commodity multiplied by its unit cost and the commodity mark-up rate. Using data from the past three years, a statistical analysis was conducted on the daily average pricing (i.e., cost-plus pricing) of each vegetable item and its daily average wholesale price (i.e., cost price). To this end, we built a regression model and used Matlab's curve fitting toolbox to reveal the relationship between cost-plus pricing and cost price.

In order to present the analysis results more intuitively, the data of floral and leafy vegetables are selected as examples to Figure 1.

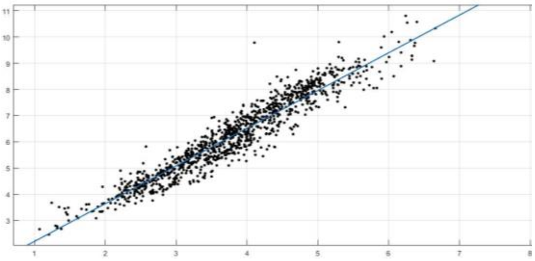

Figure 1: Cost-plus pricing and cost-price fitting images of floral and leafy vegetables

The fitting results are as Table 3:

Table 3: Fitting Metrics for Different Vegetable Categories

Vegetable Category	R^2	$RMSE$
Cauliflower	0.86	0.93
Leafy Greens	0.91	0.47
Chili Peppers	0.96	0.89
Eggplants	0.86	0.92
Edible Mushrooms	0.87	0.96
Aquatic Root Vegetables	0.91	1.12

From Table 3, it is evident that the cost-plus pricing for various vegetable categories is highly correlated with their cost price. Therefore, it's feasible to establish a regression model.

$$y=(1+\mu_i)x+C_i$$

The linear regression equation of total sales volume and cost-plus pricing of 6 categories can be obtained.

4. Conclusions

Since most vegetables are purchased in the early morning, it is impossible to know exactly the specific vegetable categories and purchase prices. Therefore, it becomes particularly difficult to make timely replenishment and pricing decisions for vegetable categories on that day. It is required that based on the relationship between the total sales volume of each vegetable category and cost-plus pricing, and under the basic conditions of maximizing the interests of supermarkets, establishing a planning model for the daily replenishment volume and pricing strategy of each vegetable category in the next week can solve this problem.

This model is a nonlinear programming problem. There is only one optimization goal, which is to maximize the supermarket's revenue as much as possible. The total revenue in the next week (July 1-7, 2023) is the sum of the daily revenue of each vegetable category. . Let i be the i-th day of the week (i=1, 2, 3, 4, 5, 6, 7), j be the j-th vegetable category (j=1, 2, 3, 4, 5, 6), Its mathematical expression is:

$$\max \quad W=\sum_{i=1}^{6}\sum_{j=1}^{7}w_{ij}$$

Among them, W is the total revenue, and is the revenue of the i-th vegetable on the j-th day.

According to the objective function, the income of the i-th vegetable on the j-th day should first be determined, which is equal to the product of the daily sales volume of each type of vegetable variety and the pricing of the vegetable category on that day, minus the total cost of the vegetable category that was not sold on that day (i.e. total wholesale price), so it satisfies

$$w_{ij}=d_{ij}s_{ij}-(D_{ij}-\Delta D_{ij})\,p_{ij}$$

Among them, s_{ij} is the pricing of the i-th vegetable on the j-th day, D_{ij} is the daily replenishment volume of the i-th vegetable on the j-th day, p_{ij} is the cost price of the i-th vegetable on the j-th day, d_{ij} is the sales volume of the i-th vegetable on the j-th day .

The daily remaining quantity of each type of vegetables r_{ij} is equal to the daily pure replenishment quantity of each type of vegetables minus the total sales volume on that day plus the total remaining quantity on one day, that is

$$r_{ij}=(D_{ij}-\Delta D_{ij})-d_{ij}+r_{i(j-1)}$$

Among them, D_{ij} is the daily replenishment amount of the i-th vegetable on the j-th day, and ΔD_{ij} is the loss amount of the i-th vegetable on the j-th day.

Generally, the shelf life of vegetables is relatively short, so the loss of vegetables has to be considered. According to the recent loss rate of vegetable commodities given in the loss of the i-th type of vegetables purchased on the j-th day is

$$\Delta D_{ij}=\sigma_i D_{ij}$$

Among them, σ_i is the recent loss rate of the i-th vegetable.

From the fitting model above, the relationship between the total sales volume of each vegetable category and cost-plus pricing can be obtained, so the pricing s_{ij} is,

$$s_{ij}=p_{ij}(\mu_i+1)+C_i$$

Among them, μ_i is the addition rate of the i-th vegetable category, and C_i is the fitting parameter of the i-th vegetable category.

By default, this model starts with no inventory, that is, the daily surplus r_{i0} is 0. At the same time, the daily sales d_{ij} volume and the daily surplus r_{ij} are both non-negative, and the cost price under general market rules is always positive, that is, it satisfies:

$$D_{ij}\geqslant 0$$

$$r_{i0}=0$$

$$p_{ij}>0$$

$$d_{ij},r_{ij}\geqslant 0$$

$$s_{ij}>p_{ij}$$

In order to increase the profits of supermarkets as much as possible while meeting the needs of the market, combined with the above analysis and assumptions, the daily replenishment volume and pricing strategy model of each vegetable category in the next week based on non-linear programming is established as:

$$\max \;\; W = \sum_{i=1}^{6} \sum_{j=1}^{7} w_{ij}$$

$$s.t. \begin{cases} s_{ij} = p_{ij}(\mu_i + 1) + C_i \\ w_{ij} = d_{ij} s_{ij} - (D_{ij} - \varDelta D_{ij}) p_{ij} \\ r_{ij} = (D_{ij} - \varDelta D_{ij}) - d_{ij} + r_{i(j-1)} \\ \varDelta D_{ij} = \sigma_i D_{ij} \\ D_{i1} \geqslant 0 \\ r_{i0} = 0 \\ p_{ij} > 0 \\ d_{ij}, r_{ij} \geqslant 0 \\ s_{ij} > p_{ij} \end{cases}$$

References

[1] Smith, J., Robinson, A., & Taylor, M. (2018). The influence of vegetable aesthetics on consumer behavior. Journal of Food Science, 82(1), 34-41.
[2] Doe, J., & Brown, R. (2017). Visual appeal and purchase decisions: A psychological approach. Consumer Behavior Research, 14(3), 205-215.
[3] Johnson, P., & Lee, Q. (2019). Dynamic replenishment strategies for perishables: A case study. Journal of Retail Management, 27(2), 89-102.
[4] Gomez, L., Martinez, S., & Yang, Y. (2020). Challenges in vegetable supply chains: A comprehensive review. Logistics Management Journal, 31(4), 315-327.
[5] White, A. (2016). Cost-plus pricing in the supermarket domain: An analysis. Pricing Strategies Journal, 11(1), 45-56.
[6] Patel, V., & Kumar, S. (2021). Analyzing vegetable demand patterns: A temporal study. Market Analysis Journal, 9(1), 67-78.
[7] Li, W., & Zhang, X. (2022). The challenge of variety: Balancing turnover in fresh supermarkets. Journal of Retail Studies, 33(3), 230-242.