Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 8, No. 3, 2024

Study on Pricing and Replenishment of Vegetable Commodities based on TOPSIS and ARIMA Models

Shitan Niu *

School of Management Engineering, Qingdao University of Technology, Qingdao, Shandong, China
*** Corresponding author Email:** 13053720677@163.com

Abstract: This study focuses on sales forecasting and pricing optimization of vegetable categories in fresh produce supermarkets. By analyzing the sales data, correlation analysis and clustering methods were used to reveal the correlation patterns among different vegetable categories. The ARIMA model was used to predict the daily replenishment quantity, and TOPSIS and entropy weighting methods were combined to screen out 27 high-quality individual products with high sales volume and stable price to achieve the maximum profit target. The study provides effective sales strategies and operational decision support for fresh food supermarkets, which has practical significance and reference value.

Keywords: TOPSIS; ARIMA Model; K-means Clustering.

1. Introduction

With the growing pursuit of healthy lifestyles by consumers, fresh produce supermarkets, as the main venue for vegetable sales, are facing increasing challenges and opportunities [1]. The purpose of this paper is to explore how to optimize the sales strategy of vegetable categories to cope with the challenges brought by the short freshness period and to improve sales efficiency and profit. In fresh food supermarkets, there are a variety of vegetable categories, and there is a complex dynamic relationship between sales and supply, which requires merchants to accurately predict the sales volume and flexibly adjust the replenishment strategy [2].

By analyzing and modeling the sales data, this study will apply Spearman correlation coefficient, cluster analysis and ARIMA model to explore the correlation, trend, and prediction of sales volume of categories. At the same time, TOPSIS and entropy weight method will be combined to screen out high-quality individual products with high sales volume and stable price, so as to realize the goal of maximum profit. These research results will provide effective sales decision support for fresh food supermarkets and help merchants better meet consumer demand, enhance competitiveness, and realize sustainable operation. Through the discussion of this study, it is expected to provide more scientific management ideas and methods for fresh food supermarkets from sales to purchasing.

2. Data Preprocessing

First of all, data preprocessing, because the sales volume of the category is a continuous variable that changes over time, choose Spearman correlation coefficient for correlation analysis, in order to verify the correlation law between categories, use k-means clustering method for each category of vegetables, and at the same time select a representative single product for correlation analysis and cluster analysis, to prove that the correlation between the categories will be affected by the influence of different single product leads to changes in correlation, and there is also a correlation between different single products. It is proved that the correlation between categories will be influenced by different single products, and there is also correlation between different single products.

2.1. Correlation Analysis

Considering that category sales are continuous variables that change over time, Pearson's correlation coefficient and Spearman's correlation coefficient are considered for correlation analysis.

The Pearson correlation coefficient measures the linear relationship between two variables and may not reflect the relationship if it is non-linear [3]. This means that Pearson correlation coefficient may underestimate or overestimate the association between variables. In addition to this, the Pearson correlation coefficient is sensitive to the skewness of the data. If the data distribution is very skewed, Pearson correlation coefficient may not be a good choice; while Spearman correlation coefficient is more general, it is based on the rank order instead of the original value, so it can capture the nonlinear relationship, and Spearman correlation coefficient is robust, even if there are more duplicates in the data, Spearman correlation coefficient can still work effectively [4]. Spearman correlation coefficient is robust, even if there are many repetitive values in the data, Spearman correlation coefficient can work effectively. Therefore, Spearman correlation coefficient was chosen to analyze the correlation of six categories of vegetables according to the sales volume of different categories.

Spearman correlation coefficient is mainly used to measure the correlation between X and Y variables. The correlation between two variables can be presented by a monotonic function, so that the correlation coefficient ρ between two variables is satisfied at +1∼-1. It is assumed that the number of elements of the two random variables X, Y are N, and when the $i-th$ value is taken, it is denoted by X_i, $Y_i$respectively. Sorting the two random variables X and Y, we obtain the ordered set x and y of the elements of the two random variables. The elements x_i and y_i are the ordering of X_i in X and Y_i in Y. The formula of Spearman's correlation coefficient is as follows. The result is shown in Table.1.

$$\rho = 1 - \frac{6\sum_{i=1}^{n}(x_i - y_i)^2}{N(N^2-1)} (1 \le i \le N) \quad (1)$$

Generally, $|\rho| > 0.95$ indicates significant correlation; $|\rho| \geqslant 0.8$ indicates high correlation; $0.5 \leqslant |\rho| < 0.8$ moderate correlation; $0.3 \leqslant |\rho| < 0.5$ low correlation; and $|\rho| < 0.3$ very weak correlation, which is considered as no correlation.

Table 1. Correlation coefficients between categories

	Foliage	Mushroom	Aquatic root	Cabbage	Pepper	Eggplant
Foliage	1 ***	0.502 ***	0.301 ***	0.5 ***	0.571 ***	0.471 ***
Mushroom	0.502 ***	1 ***	0.626 ***	0.329 ***	0.3 ***	−0.099 **
Aquatic root	0.301 ***	0.626 ***	1 ***	0.064	0.08 *	−0.155 ***
Cabbage	0.5 ***	0.329 ***	0.064	1 ***	0.482 ***	0.347 ***
Pepper	0.571 ***	0.3 ***	0.08 *	0.482 ***	1 ****	0.511 ***
Eggplant	0.471 ***	−0.099 **	−0.155 ***	0.347 ***	0.511 ***	1 ***

Note: ***, **, * represent 1%, 5%, 10% level of significance respectively.

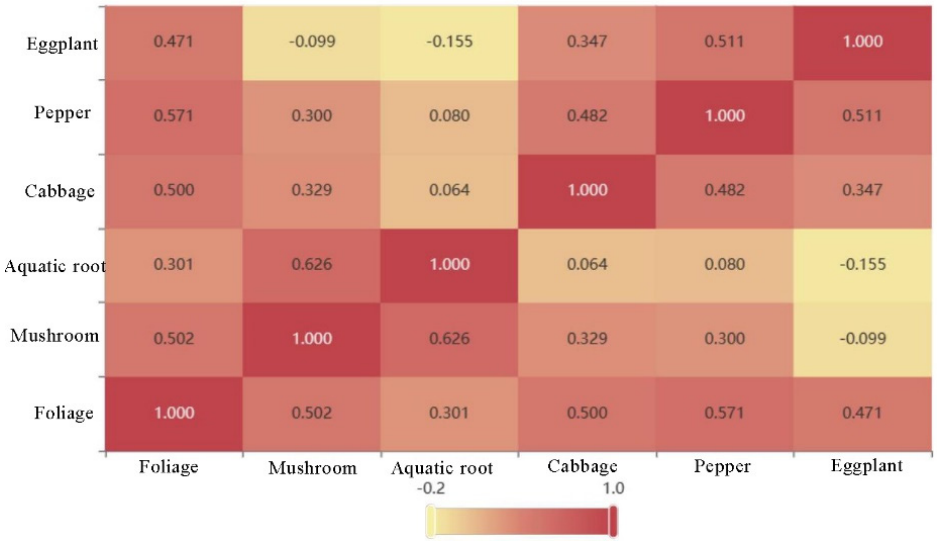

Figure 1. Heat map of correlation

The Spearman correlation coefficients were obtained by SPSS and heat map was plotted, as shown in Figure 1. It was found that there was a high correlation between foliage and mushroom, cabbage and pepper, moderate correlation between mushroom and aquatic root and foliage, and a high correlation between pepper and eggplant. And there is a negative correlation between eggplant and aquatic root and mushroom. We then considered whether people would avoid these negatively correlated categories when purchasing food, and whether they would choose to purchase items with high correlation together.

2.2. Cluster Analysis

To verify the correlation pattern between categories, we categorize the data by the sales volume of the categories as a criterion and use the K-means clustering method for each category of vegetables, we can discover the intrinsic data patterns and similarities, to deepen the insight into the structure of the data [5]. Using the law of time to determine the optimal number of clusters k in the K-means clustering algorithm, to get the k value of K-means is 3, using SPSS software, according to the daily sales volume of the category of six kinds of vegetables category clustering, get the following results in Table. 2 and Figure 2.

Table 2. Coordinates of clustering center points of each category

Clustering	Pepper	Eggplant	Mushroom	Cabbage	Aquatic root	Foliage
1	117.211	22.567	99.729	52.076	53.874	255.493
2	438.710	67.814	392.921	128.277	208.922	635.112
3	62.137	19.221	49.989	29.850	26.280	137.085

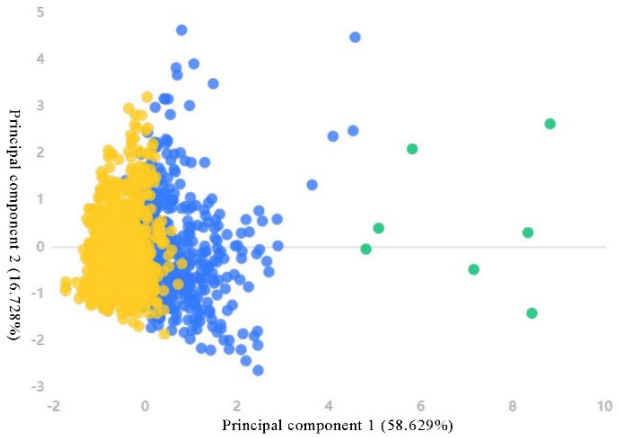

Figure 2. Clustering scatter plot

The frequency of clustering category 1 is 391 with a percentage of 36.037%, the frequency of clustering category 2 is 7 with a percentage of 0.645%, and the frequency of clustering category 3 is 687 with a percentage of 63.318%.

Table. 3. Evaluation indicators

Contour coefficients	DBI	CH
0.406	0.873	658.309

To verify the accuracy of the clustering results, a contour coefficient evaluation metric was introduced. The contour coefficient combines intra-cluster tightness and inter-cluster separation and is used to measure the degree of fitness of each sample in the clustering. As shown in Table. 3, the average

contour coefficient of this classification is greater than 0 and close to 1, the DBI value is small, and the CH value is large, so its clustering effect is better. The clustering results were found to be the same as the analysis, and there is a problem of matching when people buy. However, considering that a category contains a variety of items, the correlation between different categories may increase or decrease due to different items, and we also find that the negative correlation is not strong in the heat map, so we consider selecting representative items for correlation analysis and clustering analysis to further verify the idea.

Regarding the representative products, the first product selected for correlation analysis was bubble pepper (fine) in the pepper category and sweet potato tips in the foliage category by Spearman's coefficient. K-mean cluster analysis was conducted to analyze the sales volume of the two items.

Table 4. Spearman's correlation coefficients of the two types of products

	Sweet potato tips	Bubble pepper (fine)
Sweet potato tips	$1(0.000^{***})$	$-0.013(0.665)$

Note: $***$, $**$, $*$ represent 1%, 5%, 10% level of significance respectively.

As shown in Table. 4, it is found that although there is a positive correlation between the two vegetable categories, there is a negative correlation between the items contained in the two categories, which indicates that the correlation between the categories does not determine the correlation between the items, i.e., the items have their own correlation with each other.

3. Replenishment Forecast

Calculate the total amount of replenishment for each day of the coming week. The total amount of replenishment can usually be analyzed based on the historical sales of the product, and since the required replenishment time is relatively short, the known historical sales in 2021 and 2022 at the same period are used to forecast the replenishment amount.

ARIMA is a statistical method used for time series analysis and forecasting [6], in ARIMA(p, d, q) model, AR stands for the autoregressive process, p is the number of autoregressive terms; MA stands for the moving average process, q is the number of moving average terms; d stands for the number of times of differentiation needed to make the series smooth, the ARIMA model was created by the ARIMA method through SPSS for the forecasting of the time series of the six kinds of vegetables, and the results were analyzed. The model was created using the ARIMA method for time series forecasting of six vegetable categories using SPSS, and the results were analyzed.

Table 5. Model statistics

Model	Number of predictor variables	Model fit statistics		Ljung-BoxQ(18)			Outliers
		Outlier	R^2	Statistic	DF	Significant	
Mushroom	1	0.089	0.089	15.710	18	0.613	0
Pepper	1	0.000	0.001	14.550	18	0.693	0
Aquatic root	1	0.005	0.196	8.561	18	0.969	0
Eggplant	1	0.433	0.433	16.406	17	0.495	0
Foliage	1	0.001	0.039	27.243	18	0.075	0
Cabbage	1	0.004	-0.921	16.499	18	0.558	0

Table. 5 shows the fitting statistics of ARIMA model for 6 types of vegetables. In the statistics of "Ljung-Box Q (18)", the significance of the six materials were 0.613, 0.692, 0.969, 0.495, 0.075 and 0.558, which were all greater than 0.01 and even greater than 0.1, and the original hypothesis was accepted, and the residuals of the six materials conformed to the random distribution. At the same time, the outliers are all 0, which means that the fitted data of the six vegetables are acceptable. Therefore, the ARIMA model was used to predict the daily replenishment quantity from July 1 to 7, 2023, and the following results in Table. 6 were obtained.

Table 6. Statistics of predicted values

Date	2023.7 .1	2023.7 .2	2023.7 .3	2023.7 .4	2023.7 .5	2023.7 .6	2023.7 .7
Mushroom	37.1473	36.5918	36.0363	35.4809	34.9254	34.3700	33.8144
Cabbage	43.3341	47.4128	46.6050	47.9506	48.4628	49.4088	50.2510
Pepper	38.5395	37.2533	36.0155	34.8260	33.6848	32.5920	31.5476
Aquatic root	46.8654	50.8446	55.0037	59.3427	63.8615	68.5602	73.4388
Eggplant	20.4104	18.8130	17.6026	16.3922	15.1817	13.9713	12.7609
Foliage	128.6172	123.0134	117.2326	111.2747	105.1398	98.8279	92.3389

4. Restocking Pricing Strategy

4.1. Selection of Available Items

The following points are taken into consideration in the revenue evaluation, first, 27-33 items are selected for the replenishment program; second, the average sales of these 27-33 groups of items between June 24 and 30 are greater than 2.5 kg, and the profit is maximized.

First of all, we select the sales volume of all single products between June 24-30, between June 24-30, the daily sales volume of many single products is 0, so the deletion operation is carried out, after the deletion of the end of the only 44 groups of data left, as shown in Table. 7. Through the use of the weighting method by the sales volume of these six days for the weighting, the use of TOPSIS superiority and inferiority solution distance method to solve the evaluation of the object of the superiority and inferiority of the solution of the distance, sorted by the high that is, the overall score index is high that is, six days of the sales of the change is more stable, the future value of the prediction of the more accurate [7].

Table 7. TOPSIS method solved 44 kinds of single product

Single product	D+	D-	C	Sorting results
Yunnan lettuce	1.605	14.346	0.899	1
Millet pepper	9.27	7.225	0.438	2
Golden needle mushroom	8.761	6.429	0.423	3
Screw pepper	9.472	5.275	0.358	4
Wuhu green pepper	9.649	5.181	0.349	5
Purple eggplant	10.289	4.651	0.311	6
... ...	... ...	... .	... .	... ...
Wood ear vegetable	14.305	0.257	0.018	41
White jade mushroom	14.364	0.232	0.016	42
Green thread pepper	14.403	0.161	0.011	43
Purple eggplant	14.396	0.131	0.009	44

Because only consider the 24-30 seven-day sales volume, and did not fully take into account the return situation and the discount situation and the sales volume on July 1, so first of all the 44 sets of data for the predicted value of the solution, because the problem for seven days of data to predict the predicted value of a day of small samples, so we consider the use of linear regression, exponential smoothing and other methods of prediction, taking into account the need to analyze the impact of recent multiple time points on July 1, the three exponential smoothing method of prediction of short-term data is applicable, so we consider the exponential smoothing method. Considering the need to analyze the impact of many recent time points on July 1, the three times exponential smoothing method is suitable for predicting the short-term data, so we consider the exponential smoothing method.

$$L(t) = \alpha * Y(t) + (1-\alpha) * [L(t-1) + T(t-1)] \quad (2)$$

$L(t)$ denotes the smoothing level (mean) at time t. a is the smoothing coefficient, usually between 0 and 1, to control the weight of the new observations. a is the smoothing coefficient, usually between 0 and 1, which controls the weight of the new observation. $Y(t)$ is the observed value at time t. $L(t-1)$ is the observed value at time t. $L(t-1)$ is the smoothing level at time $t-1$. $T(t-1)$ is the trend at time $t-1$.

$$T(t) = \beta * [L(t) - L(t-1)] + (1-\beta) * T(t-1) \quad (3)$$

$T(t)$ denotes the trend at time t. β is the trend smoothing coefficient, usually also between 0 and 1. β is the trend smoothing coefficient, usually also between 0 and 1.

$$S(t) = \gamma * [Y(t) - L(t)] + (1-\gamma) * S(t-m) \quad (4)$$

$S(t)$ denotes the seasonal component at time t. γ is the seasonal smoothing coefficient. γ is the seasonal smoothing coefficient, usually also between 0 and 1. m denotes the seasonal period, e.g., if the season repeats annually, m may be 12 (months). The final predicted values are.

$$F(t+h) = L(t) + h * T(t) + S(t-m+h) \quad (5)$$

Through the prediction, we found that the predicted values of five items, baby vegetables, purple eggplant, green eggplant, colorful peppers, and enoki mushrooms, were negative, which had a side effect on the prediction of the maximal return of the superstore, and so they were deleted. In addition to the returns of June 8 and the corresponding wear and tear rate in June, the discount rate from 24th to 30th is also considered to select the 27 items that can be sold in June.

4.2. Developing Replenishment and Pricing Strategies for Individual Items

After predicting the sales volume of these 27 items, we need to analyze and solve the pricing strategy, analyzing the pricing strategy, we also need to consider the previous unit price and wholesale price of each item, as well as also need to take into account the average sales volume of each item ≥2.5, so we get the dynamic programming equation: ≥2.5.

$$\max L = V \times [S \times (1-\beta) - C] \quad (6)$$

$$\text{s.t.} \begin{cases} S \le S_{\max} \\ V_d \ge 2.5 \\ W_{\min} \le C \le W_{\max} \end{cases} \quad (7)$$

$$\pi = L \times S_n \quad (8)$$

L is the profit, V is the daily sales volume of each item, β is the daily wastage rate of each item, $S_{\max}$ is the highest historical unit price of the item for three years, W_{min} and W_{max} are the lowest and highest historical unit prices of the item for three years.

According to the triple exponential smoothing method, calculate the basket selected 27 of the sellable single product sales volume on July 1, at the same time according to the dynamic programming equation, the use of SPSS to solve for each single product of the maximum profit, that is, the unit price of the sale, and finally arrive at the July 1, the maximum revenue of each single product, as shown in Table. 8.

Table 8. Predicted sales volume and profit of individual products on July 1

Single product category	Sales volume forecast	Sales unit price	Maximum revenue
Broccoli	18.730	19.48	236.616
Zhijiang green stalks loose flower	14.889	13.89	43.756
Spinach	10.837	6.77	60.939
Sweet potato tips	7.135	19.80	105.052
Wood ear vegetable	4.426	19.66	115.776
Milk cabbage	11.209	22.13	163.089
... ...	... ...	... ...	... ...

5. Conclusion

Based on the research on the sales strategy of vegetable categories in fresh supermarkets, we draw the following conclusions: first, through correlation analysis and cluster analysis, it is found that there is a certain degree of correlation between different vegetable categories, and that the influence of the single category will lead to changes in correlation. Second, using ARIMA model to predict daily replenishment can help supermarkets better adjust replenishment strategies and improve sales efficiency. Finally, through TOPSIS and entropy weight method to screen out the high-quality single products with high sales volume and stable price, combined with the return situation and the loss rate, the optimal 27

sellable single products are finally determined, to realize the goal of maximum profit. This study provides a scientific basis and methodology for sales decision-making in fresh food supermarkets, helping merchants to better cope with the challenges posed by short freshness period, enhance competitiveness and realize sustainable operation.

References

[1] Augustin M A, Sanguansri L, Fox E M, et al. Recovery of wasted fruit and vegetables for improving sustainable diets[J]. Trends in Food Science & Technology, 2020, 95: 75-85.

[2] Priore P, Ponte B, Rosillo R, et al. Applying machine learning to the dynamic selection of replenishment policies in fast-changing supply chain environments[J]. International Journal of Production Research, 2019, 57(11): 3663-3677.

[3] Armstrong R A. Should Pearson's correlation coefficient be avoided? [J]. Ophthalmic and Physiological Optics, 2019, 39 (5): 316-327.

[4] Alsaqr A M. Remarks on the use of Pearson's and Spearman's correlation coefficients in assessing relationships in ophthalmic data[J]. African Vision and Eye Health, 2021, 80(1): 10.

[5] Maugeri A, Barchitta M, Favara G, et al. The application of clustering on principal components for nutritional epidemiology: A workflow to derive dietary patterns[J]. Nutrients, 2022, 15(1): 195.

[6] 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.

[7] Tian J, Wu Z, Qu X. Research on Fire Alarm System Based on Entropy Power Method and Topsis Superiority and Inferiority Solution Distance Method[J]. International Journal of New Developments in Engineering and Society, 2023, 7(2).