

Research on Vegetable Sales Data Analysis and Cost Pricing Modeling

Yikang Li¹, Hang Zhao², Yanran Wang³, Bohao Yu^{1,*}

¹ School of Mathematics, Yunnan Normal University, Kunming, China, 650500

² School of Life Science, Yunnan Normal University, Kunming, China, 650500

³ Faculty of Chemical Engineering, Kunming University of Science and Technology, Kunming, China, 650500

* Corresponding Author Email: yubohaobaihe@126.com

Abstract. Vegetable commodities have a shorter freshness period, many varieties, and a special time of purchase, so it is important to comprehensively analyze the automatic replenishment and pricing strategy of vegetable commodities for the profitability of fresh food superstores. First of all, in the correlation of vegetable categories and single product sales volume, Pearson correlation coefficient analysis is used, to obtain the strong correlation between chili peppers and foliage categories and various vegetable categories. Then based on correlation the best clustering results were obtained by K-means clustering analysis divided into three main categories, namely solanula; flowers and leaves and cauliflower; peppers, aquatic rhizomes, and edible fungi. Next using entropy weights TOPSIS method is used to give the weights of four indicators: total sales, average unit price, average wholesale price, and average wastage rate of each category of vegetable commodities for seven days, and the top 27 vegetable items are sorted to get the best sellable vegetables. Based on the above ranking of the individual vegetable items, Excel was used to forecast the cost pricing of the vegetable commodities to get the cost pricing of the 27 individual vegetable items for the next seven days.

Keywords: Pearson correlation coefficient, entropy weight TOPSIS, cost pricing prediction.

1. Introduction

In today's fresh produce superstores, vegetables have a short shelf life and are traded at special times, and superstores will sell vegetables at a discount for shipping damage and deterioration. As a result, merchants are often required to restock without being sure of the purchase price and individual products. Thus, a reliable market demand analysis plays a crucial role in replenishment and pricing decisions.

Currently, Pearson correlation analysis and entropy weight TOPSIS method are widely used in commodity analysis. Xu used the Pearson correlation coefficient has been used in the field of clothing consumption [1]. Wang et al. utilized the entropy weight-TOPSIS method to comprehensively evaluate the fruit quality of three new varieties of roses for fruit use, which was applied in the field of fruits and vegetables [2]. Quanjie Shen and Yingnan Liang conducted an optimisation study on the model of cold chain supply of agricultural products logistics in China based on the Internet of Things (IoT) technology. [3] Hongshen Yu et al. Agricultural product logistics efficiency analysis and optimisation study with actual data from Jilin province. [4] Zheng Yuting et al. study the optimal (maximum marginal profit) decision of distributors when preservation effort is exogenous versus when order quantity is exogenous by combining data from actual firms. [5]

In this study, Pearson's correlation coefficient was first used to study the correlation among six vegetable categories, and then K-means cluster analysis was applied to divide the vegetable categories as a whole. Finally, the entropy weight TOPSIS method was used to calculate the weights of the four indicators, and the top 27 vegetable items and the top 27 vegetable items were ranked.

2. Correlation analysis of vegetable sales data

2.1. Study of Pearson's correlation coefficient

Pearson, Kendall, and Spearman are used to measure the correlation between data in statistics. In this study, the Pearson correlation coefficient is used to explore the correlation between vegetable category and sales volume and to measure the degree of linear correlation between data. Pearson correlation coefficient was first designed by Karl Pearson as a statistical index to study the degree of linear correlation between variables and is measured from -1 to +1. Pearson correlation coefficient is a statistical indicator designed by Karl Pearson to study the degree of linear correlation between variables and is measured from -1 to +1, reflecting the direction and degree of the trend of change between the two variables, with 0 indicating that the two variables are not correlated, a positive value indicating positive correlation, a negative value indicating negative correlation, and a larger value indicating a stronger correlation [6]. Pearson correlation coefficient, based on the similarity represented by Euclidean distance, is centralized to the computed values and then seeks the centered result of the cosine distance. The difference between the different measures of the variables can be eliminated in the above calculation process with the following formula [7]: (Data source: <http://www.mcm.edu.cn/>)

$$\rho(X, Y) = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sqrt{\sum_{i=1}^n (X_i - \mu_X)^2} \sqrt{\sum_{i=1}^n (Y_i - \mu_Y)^2}} \quad (1)$$

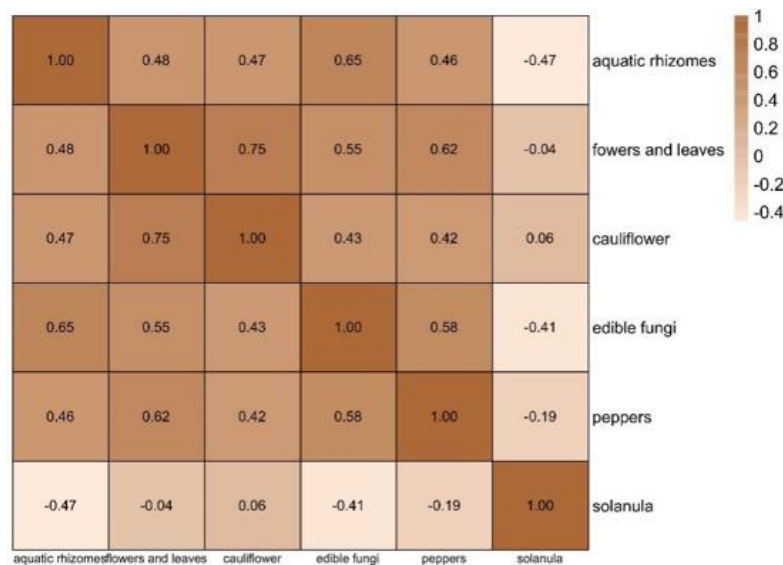


Figure 1. Pearson correlation analysis heat map.

From the heat map of Pearson correlation analysis, it is easy to see that the correlation of some vegetable categories is not high, for example, eggplant and cauliflower, edible mushrooms and aquatic roots and tubers, etc., and their low correlation may be related to the fact that all vegetable categories are main dishes. Some vegetable categories have a high correlation, and the correlation between chili peppers and cauliflower and various vegetable categories is high, which may be because one of them is a main dish and the other is a side dish, which needs to be purchased together.

2.2. K-means cluster analysis of vegetable categories

In this study, K-means clustering combined with analysis was used to investigate the correlation of vegetable categories. Firstly, the sum of error squares needs to be judged to determine the best clustering, and the sum of error squares formula is as follows:

$$H = \sum_{i=1}^K \sum_{\alpha \in c_i} |d - m_i|^2 \quad (2)$$

Where H is the sum of error squares, ci denotes the i th class, d is the sample point of ci , and mi is the center of mass of ci .

After determining the optimal number of clusters, cluster analysis is performed, assuming that there are D clustered objects, each consisting of n elements, which are first normalized:

$$t'_{ij} = \frac{t_{ij} - \bar{t}_j}{s_j} (i=1, 2, \dots, m, j=1, 2, \dots, n) \quad (3)$$

$$\bar{t}_j = \frac{1}{m} \sum_{i=1}^m t_{ij} \quad (4)$$

$$s_j = \sqrt{\frac{1}{m} \sum_{i=1}^m (t_{ij} - \bar{t}_j)^2} \quad (5)$$

Where denotes the standard deviation, denotes the arithmetic mean of the j th element.

The arithmetic mean of each element obtained after the standardization process is 0, and the standard deviation is 1. K initial clustering centers are determined based on the optimal number of clusters, and then the samples are assigned to the closest clusters according to the principle of the shortest Euclidean distance. The Euclidean distance Q from the clustering object i to the initial clustering center is:

$$Q = \sqrt{\sum_{i=1}^m (T_i - G)^2} \quad (6)$$

Assume that at this time one of the initial clustering center coordinates is G , and the coordinates of the i th clustering object are T_i . Finally, repeat the above steps with each sample average as the new clustering center, and finally get the best clustering results [8].

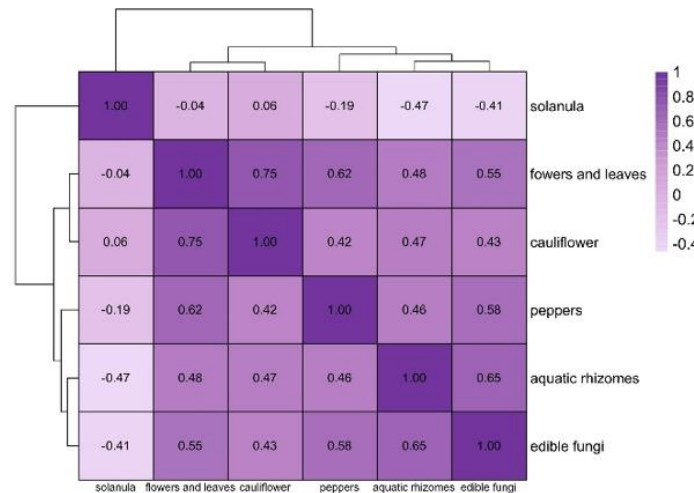


Figure 2. Correlation Analysis Cluster Plot.

Through the combined study of correlation and clustering, the vegetable category as a whole can be divided into three major categories, namely solanula; flowers and leaves and cauliflower; peppers, aquatic rhizomes, and edible fungi. For customers, every time they buy vegetables, they may for customers, each time they buy vegetables, they may not only buy one type of vegetable, but they may buy several types of vegetables together.

3. Sorting and Cost Pricing of Vegetable Commodities

3.1. Entropy weight TOPSIS method to synthesize the evaluation of each vegetables

The entropy weighting method is a method to obtain the discrete degree of each indicator based on the information entropy of each indicator, to objectively obtain the weights of each indicator,

effectively avoiding subjectivity when assigning the weights, and having a certain degree of accuracy compared with the subjective method such as hierarchical analysis. Information entropy represents the degree of disorder of the system, the more disordered the system is, the smaller the information entropy of the indicator is, the greater the degree of discretization, the greater the amount of effective information that can be provided, and the higher the weight of the indicator in the comprehensive evaluation [9].

Through the study on the sale of vegetables, the researchers selected the four indicators of vegetable single products the total sales of each category for seven days, the average sales unit price, the average wholesale price, the average loss rate, and the four indicators are standardized.

Step 1 Assuming that the number of comprehensive evaluation objects x of education development is m , the number of comprehensive evaluation indicators of education development for each x is n , and x_{ij} is the raw data of the j th indicator of the i th evaluation object, the original matrix X of comprehensive evaluation is:

$$X = (x_{ij})_{m \times n}, i = 1, 2, \dots, m, j = 1, 2, \dots, n \quad (7)$$

Step 2 since the scale of each indicator is different, it is necessary to standardize the data of each indicator in the original matrix X . For positive indicators, the standardization formula is:

$$Y = (y_{ij})_{m \times n} = \frac{x_{ij} - x_{j\min}}{x_{j\max} - x_{j\min}} \quad (8)$$

For negative indicators, the normalization formula is:

$$Y = (y_{ij})_{m \times n} = \frac{x_{j\max} - x_{ij}}{x_{j\max} - x_{j\min}} \quad (9)$$

Where: Y is the standardized comprehensive evaluation matrix whose elements are y_{ij} : x_j , $\min$, x_j , and $\max$ are the minimum and maximum values of the j th evaluation index of x , respectively.

Step 3 Calculate the information entropy e_j of indicator j , using Y .

$$e_j = -k \sum_{i=1}^m p_{ij} \ln p_{ij} \quad (10)$$

Included among these $p_{ij} = \frac{y_{ij}}{\sum_{i=1}^m y_{ij}}, k = \frac{1}{\ln m}$, when $p_{ij}=0$, make $\lim_{p_{ij} \rightarrow 0} p_{ij} \ln p_{ij} = 0$

Step 4 Based on the value of the information entropy of the indicator, define the weight w_j of the indicator j as:

$$w_j = \frac{(1 - e_j)}{\sum_{j=1}^n (1 - e_j)}, w_j \in [0, 1], \sum_{j=1}^n w_j = 1 \quad (11)$$

Step 5 Use the weights calculated in (5) to obtain the weighting matrix R .

$$R = (r_{ij})_{m \times n}, r_{ij} = w_j y_{ij} (i = 1, 2, \dots, m; j = 1, 2, \dots, n) \quad (12)$$

Step 6 Determine the positive ideal solution for each indicator as $S_j^+ = \max\{r_{1j}, r_{2j}, \dots, r_{mj}\}$. The negative ideal solution is $S_j^- = \min\{r_{1j}, r_{2j}, \dots, r_{mj}\}$

Step 7 Calculate the weights of each evaluated object and the Euclidean example distances of positive and negative ideals:

$$\begin{cases} D_i^+ = \sqrt{\sum_{j=1}^n (s_j^+ - r_{ij})^2} \\ D_i^- = \sqrt{\sum_{j=1}^n (s_j^- - r_{ij})^2} \end{cases} \quad (13)$$

Step 8 Calculate the closeness of the evaluation object to the ideal C_i .

$$C_{ij} = \frac{D_i^-}{(D_i^+ + D_i^-)} \quad (14)$$

C_i takes a value in the range of [0, 1], the larger the C_i the higher the overall evaluation [10].

Table 1. Four indicator weights

Item	information entropy value e	Information utility valued	Weighting factor w
Sum:Sales(kg)	0.8813	0.1187	35.37%
Average:Unit price(yuan/kg)	0.9208	0.0792	23.60%
Average:Wholesale Price(yuan/kg)	0.9032	0.0968	28.85%
Average:Wastage rate(%)	0.9591	0.0409	12.18%

Table 2. Top 27 Vegetable Singles Sorting.

Term	Positive ideal solution distance D	Negative ideal solution distance D-	Relative proximity C	Ordering results
Honghu Lotus Root	0.317	0.356	0.529	1
Scallops				
Yunnan Lettuce(portion)	0.337	0.359	0.516	2
Xixia Flower Mushroom	0.317	0.335	0.514	3
Wild Powdered Lotus Root	0.36	0.352	0.494	4
Red Pepper	0.361	0.265	0.424	5
Colorful Pepper	0.374	0.256	0.407	6
High Melon	0.363	0.244	0.402	7
High Melon	0.378	0.252	0.4	8
Millet Pepper(portion)	0.368	0.241	0.396	9
Broccoli	0.321	0.208	0.393	10
Yunnan Oil Wheat Vegetable(portion)	0.368	0.239	0.393	11
Lotus Root	0.349	0.215	0.381	12
Spinach	0.394	0.202	0.339	13
Field Chrysanthemum	0.406	0.193	0.322	14
Rhododendron	0.396	0.186	0.319	15
Zijiang Green Stem	0.384	0.178	0.317	16
Scattered Flowers				
Screw Pepper	0.372	0.165	0.307	17
Wuhu Green Pepper	0.393	0.164	0.295	18
Golden Needle				
Mushroom(box)	0.428	0.176	0.291	19
Bamboo Leaf Lettuce	0.41	0.158	0.277	20
Long Thread Eggplant	0.399	0.149	0.272	21
Baby Cabbage	0.399	0.136	0.255	22
Screw Pepper(portion)	0.409	0.137	0.251	23
Purple Eggplant	0.405	0.134	0.249	24
Small wrinkled skin (portion)	0.44	0.129	0.227	25
Amaranth	0.434	0.126	0.225	26
Yunnan Lettuce	0.435	0.124	0.222	27

The above Table 2 shows that the total seven-day sales volume of each category has the largest weight in the indicator weights. The analysis is based on the four indicators in Table 1, ranked

according to the relative proximity size of each category, we selected the top 27 as the best available July 1, 2023 products.

3.2. Cost projections of individual vegetables

For the above 27 categories of commodities, according to meet the universal laws of economics as well as commodities that do not meet the universal laws of economics, the commodities are divided into 2 categories for prediction, which will meet the universal laws of economics using exponential smoothing method, the use of Excel worksheet prediction for prediction, and the remaining use of the average value prediction. See Table 3 below.

Table 3. July 1 forecasts for items with more pronounced changes in cost pricing and sales volume.

	Forecasted sales volume (in kilograms)	Forecasted pricing (in yuan)	Forecasted wholesale price (in yuan)	Profit (in yuan)
Colorful Peppers	1.51	19.61	11.92	10.49
Tall Melon	3.79	12	11.77	0.62
Tall Melon	1.21	18.2	13.48	5.17
Broccoli	19.07	10.21	7.4	48.62
Zijiang Green Stem	14.68	11.25	7.62	48.27
Scattered Flowers	8.27	12.71	9.78	23.98
Screw Pepper	11.33	6.77	4.52	23.09
Screw Pepper(portion)	5.31	24.46	18	26.05
Honghu Lotus Root				
Scallop				

Most of the cost pricing does not change with the change in sales volume by using an excel pivot table to analyze the items that do not change significantly over the seven days. See Table 4 below.

Table 4. July 1 projections for items with insignificant changes in cost pricing and sales volumes.

	Forecasted sales volume (in kilograms)	Forecasted pricing (in yuan)	Forecasted wholesale price (in yuan)	Profit (in yuan)
Spinach	0.4	11.9	9.64	0.91
Red Pepper	0.22	16.06	12.97	0.68
Enoki Mushroom(box)	1	1.6	1.45	0.14
Lotus Root	0.47	12.24	10.73	0.71
Rhododendron	0.53	11.9	9.17	1.45
Baby Cabbage	1	5.56	4.76	0.8
Field Chrysanthemum	0.47	10.64	8.24	1.12
Wuhu Green	0.37	4.42	3.37	0.39
Pepper(1)	0.18	20.4	15.6	0.88
Xixia Mushroom	0.52	3.23	2.33	0.47
Amaranth	1	2.2	1.52	0.67
Small wrinkled skin(portion)				
Wild Powdered Lotus Root	0.41	22.1	16.07	2.45
Yunnan Lettuce	0.32	7.82	5.74	0.67
Yunnan Oil	1	3.54	2.87	0.67
Lettuce(portion)	0.4	10.2	6.98	1.3
Long Thread Eggplant	0.43	3.21	2.33	0.38
Bamboo Leaf Lettuce	2.15	2.79	1	3.85
Millet Pepper (portion)				
Yunnan lettuce (portion)	3.58	4.64	1	13.03
Purple Eggplant	3.74	2.53	0.54	7.43

4. Conclusions

In the correlation of vegetable categories and single product sales, Pearson correlation analysis was used to obtain a strong correlation between pepper and flowers and leaves and various vegetable categories, and then combined with K-means clustering can be divided into three major categories of vegetable categories as a whole, which are solanula; flowers and leaves and cauliflower; peppers, aquatic rhizomes, and edible fungi. Peppers, aquatic rhizomes, and edible fungi.

The entropy weight TOPSIS method was used to analyze the category evaluation from June 24-30, 2023. Solve the weight Table 1 corresponding to the importance of each single product vegetable: sales volume 35.37%, sales unit price 23.60%, wholesale price 28.85%, and wastage rate 12.18%. From there, the categories with higher relative proximity are screened for sorting, and the top 27 are identified. The more important ones in Table 2 are red tiger lotus root, Yunnan lettuce, Xixia mushroom, and wild pink lotus root. Daily replenishment and cost pricing were predicted for July 1 by exponential smoothing. Table 3 and Table 4 show the seven-day average projections of sales volume and cost pricing for categories that have not changed significantly.

References

- [1] Yanqiu Xu. Research on the influence of apparel brands on apparel consumption psychology based on the Pearson coefficient [J]. Wool Textile Science and Technology, 2018, 46(03):59-62.DOI:10.19333/j.mfkj.2017050080304.
- [2] Wang L,Zhao HY,Du XC et al. Comprehensive evaluation of fruit quality of three new varieties of rose hips for fruit use based on entropy-weighted TOPSIS method[J/OL]. Chinese Materia Medica, 2023(08):1858-1864[2023-10-04].https://doi.org/10.13863/j.issn1001-4454.2023.08.003.
- [3] Quanjie S, Yingnan L. Effective Cold Chain Supply of Agricultural Products Logistics in China Based on Internet of Things [J]. Journal of Physics: Conference Series, 2021, 1972(1).
- [4] Yu H, Zhao L, Wu M. Study on the evaluation of agricultural product logistics efficiency in Jilin province[C]// Jinan Linfeng Culture Media Co., Ltd. Proceedings of 2020 2nd International Conference on Applied Machine Learning and Data Science (ICAMLDS 2020).IOP Publishing,2020:8.DOI:10.26914/c.cnkihy.2020.018850.
- [5] Zheng Yuting,Li Jianbin, Chen Zhiyuan et al. Optimal decision making for cold chain distributors under uncertain demand[J]. Journal of Management Science, 2019, 22(01):94-106.
- [6] Cheng Juanjuan. An empirical study on the relationship between research and teaching in colleges and universities--an analysis based on Pearson's correlation coefficient [J]. Science and Technology in Chinese Colleges and Universities, 2022(10):46-52.DOI:10.16209/j.cnki.cust.2022.10.016.
- [7] Zhao Y, Yu LJ, Su YX, et al. cleaning and de-duplication of grid load data based on cluster analysis and Pearson correlation coefficient method [J]. Ship Power Technology, 2023, 43(6):69-75. DOI:10.3969/j.issn.1003-4862.2023.06.016.
- [8] Liu QH, Tang FX. Optimization of site selection for faulty shared bicycle recycling center based on K-means clustering algorithm and center of gravity method [J]. Operations Research and Management, 2023, 32(07):85-91.
- [9] Feng Jingqian, Li Jiangyan, Song Jianfeng et al. Comprehensive evaluation of the quality of Hovenia quadrangularis herbs from different origins by entropy weight TOPSIS method [J]. Chinese Pharmacy, 2021, 32(11):1312-1318.
- [10] ZHANG Wei-Yi. Entropy weight TOPSIS method for comprehensive evaluation of educational development in Xinjiang [J/OL]. Journal of Jinan University (Natural Science Edition):1-8 [2023-09-27]. https://doi.org/10.13349/j.cnki.jdxbn.20230905.002.