Information Systems and Economics (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/infse.2023.041017
ISSN 2523-6407 Vol. 4 Num. 10

Research on vegetable bundling decisions based on K-means cluster analysis

Yining Zhou[*,#], Xingyu Zhou[#]

School of Government Audit, Nanjing Audit University, Nanjing, 211815, China
[*]Corresponding author: zynpye123@163.com
[#]These authors contributed equally.

Keywords: Spearman Correlation Coefficient, K-means Cluster Analysis, Contour Coefficient

Abstract: With the improvement of the quality of life, it has become a trend to buy vegetables in fresh agricultural products supermarkets. In the actual sales process, fresh produce usually increases supermarket revenue through bundling, so it is important to study the degree of association between each vegetable category and the correlation between single vegetable products for the bundling decision of supermarkets. In this paper, SPEARMAN correlation analysis and K-MEANS cluster analysis method are adopted to study the sales volume and pattern of each vegetable category and single vegetable product, and the time series analysis method is used to analyze the seasonal sales rules of single vegetable product and each vegetable category. This paper finds that the sales volume often reaches the maximum in winter. Finally, the optimal bundling decision and seasonal replenishment strategy are obtained according to the correlation between the vegetable category and each vegetable item and the maximum winter sales.

1. Introduction

The proportion of fresh agricultural products in the sales of major supermarket chains has gradually increased, becoming the core commodity of supermarket chains to attract customers. In the actual sales process, fresh agricultural products are often bundled with each other to increase supermarket revenue. Therefore, it is of great significance to study the degree of association between agricultural products for supermarkets to make bundled sales decisions. In this paper, Spearman's correlation coefficient is used to analyse the degree of association between agricultural products in order to determine the general direction of bundled sales. When analysing each vegetable item, due to the excessive number of items and the lack of data, the Spearman correlation coefficient analysis could not accurately analyse the correlation coefficient between the items, so this paper adopts the K-means cluster analysis method to classify each vegetable item into three major categories, and draws the conclusion that the vegetable items classified into the same category are more correlated with each other, which provides ideas for the hypermarket to make the decision of bundled sales.

2. Analysis of the distribution pattern of sales volume by category

2.1 Total sales by category

Firstly, each individual product is categorised according to its category. The distribution pattern of sales volume of each category is analysed with basic statistics. In this paper, according to the sales data of 2020.7.1-2023.6.30, the single product category is classified and summarised to obtain the total sales volume of each vegetable category, and each category accounts for the total sales volume as shown in Fig 1. It can be seen that the flower and leaf category has the largest sales, accounting for 42 per cent of the total sales, and the eggplant category has the smallest sales, accounting for 5 per cent of the total sales.

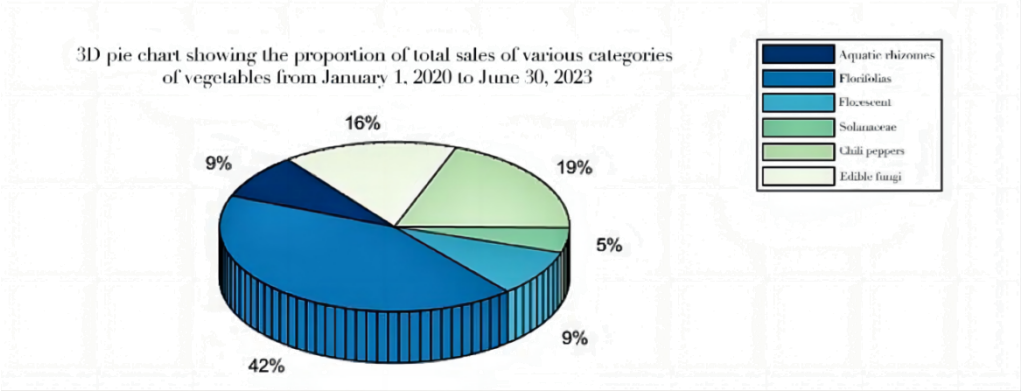
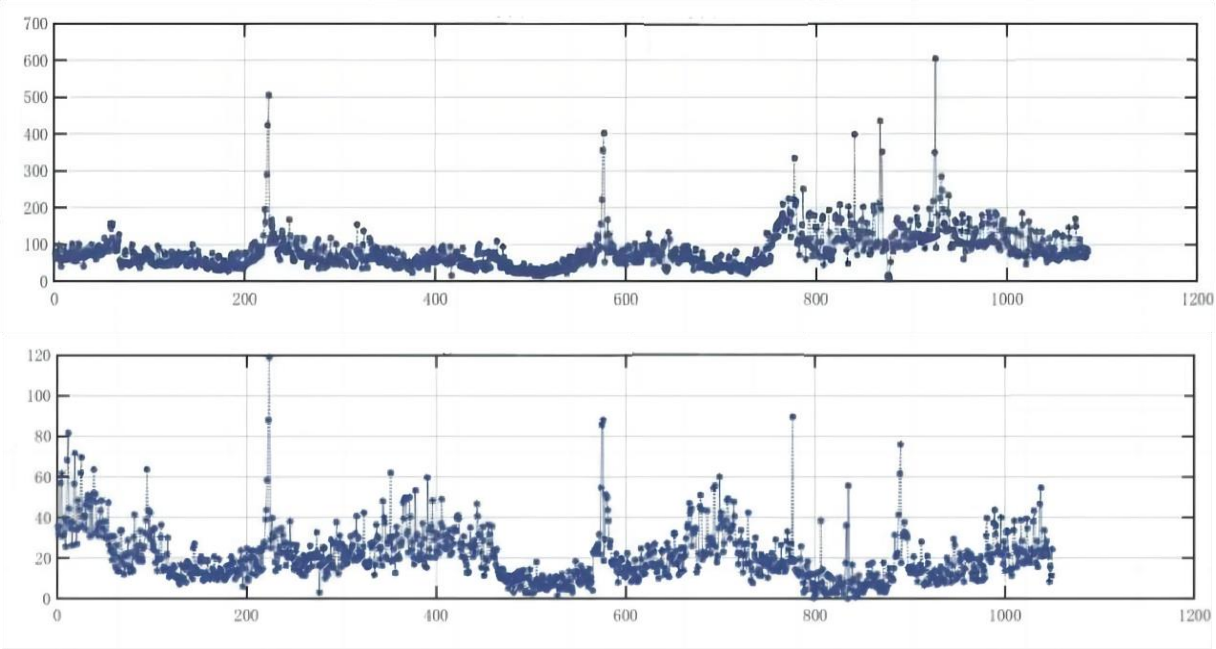


Figure 1: Three-dimensional pie chart of sales share of various types of vegetables

2.2 Time distribution by category

In this paper there is often a correlation between the sales volume of vegetable items and time[1]. So a correlation analysis is to be made between the sales quantity and time for each category. Based on the change in sales quantity of each category with time, a graph of sales quantity with time is derived.



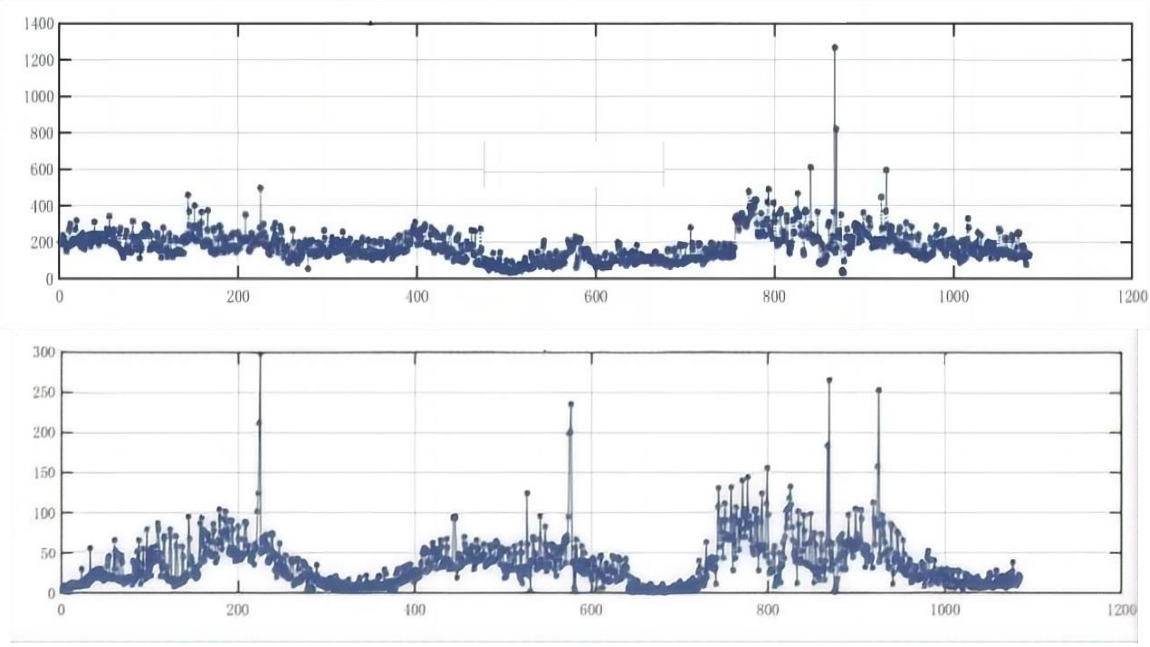


Figure 2: Sales volume by category over time

The horizontal coordinate of the Fig2 is the number of days of sales, 0 represents 1 July 2020, and the vertical coordinate represents the sales volume (kg) on that day. From the graph, it can be seen that the sales volume of each category shows an increasing and then decreasing trend from July in each year, and the sales volume reaches the maximum value in the winter time of each year, which has a strong seasonality[2] . Comparing the six categories of graphs vertically, this paper finds that sales tend to reach their maximum in winter.

3. Analysis of the distribution pattern of sales volume of each individual product

3.1 Total sales of each individual product

The total sales volume of each item was ranked and the top 10 items in terms of total sales volume are shown in Table 1.

Table 1: Total Sales Ranking

name (of a thing)	Sales volume (kg)	name (of a thing)	Sales volume (kg)
Wuhu Green Pepper(1)	28164.331	Golden Needle Mushroom(box)	15596
broccoli	27537.228	Yunnan lettuce (portion)	14325
Lotus root(1)	27149.44	Purple Eggplant(2)	13602.001
Brassica pekinensis	19187.218	Xixia Shiitake Mushroom(1)	11920.227
Yunnan lettuce	15910.461	Peppers (portions)	10833

3.2 Average daily sales volume of individual products

The daily sales of vegetables were filtered through an Excel spreadsheet to filter out the largest average daily sales of vegetables and the results are shown in Table 2.

Table 2: Table of average daily sales

name (of a thing)	Sales volume (kg)	name (of a thing)	Sales volume (kg)
Fresh Dumpling Leaves (bag) (1)	4.878	Golden Needle Mushroom (bag) (1)	1.081
Lotus seedling (pcs)	4.241	White Mushroom(box)	1.024
Fresh Dumpling Leaves (Bag) (3)	1.391	Baokang Alpine Cabbage	1.012
Brassica pekinensis	1.264	Seafood Mushroom(bag) (1)	1.008
Fresh Lotus Root Strip (Bag)	1.105	Honghu Lotus Root (Powdered Lotus Root)	1.008

4. Relationship analysis by category

In order to explore the potential correlations between the various vegetable categories, a study using correlation analysis is required. Correlation analysis is the process of analysing the degree of correlation between variables and deriving the correlation coefficient. According to the literature, there are two ways of calculating correlation coefficients, which need to be chosen according to different data types: Pearson correlation coefficient[3] is used when the data are quantitative and satisfy the normal distribution[4], and Spearman correlation coefficient[5] is used when the data are quantitative but do not satisfy the normal distribution.

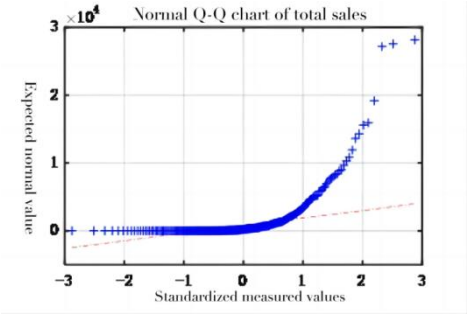


Figure 3: Normal Q-Q plot of total sales volume

As can be seen from Fig 3, the data for total sales volume was found to be not normally distributed. Therefore Spearman's correlation coefficient was used to portray the degree of correlation between the sales volume of different vegetable categories:

$$\rho = \frac{\sum_i (x_i - \bar{x})\ (y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \tag{1}$$

The heat map of correlation coefficient of sales volume data of each vegetable category is calculated, as shown in Fig. 4. Among them, the correlation coefficient of eggplant and cauliflower is larger, 0.889. In the actual sales process, merchants can bundle the vegetable categories with larger correlation coefficients to increase sales.

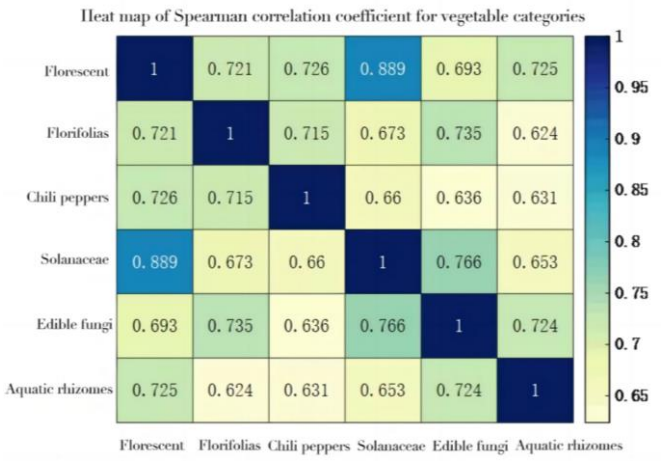


Figure 4: Heat map of Spearman's correlation coefficient for vegetable category

5. Relationship between individual products

Due to the large number of single product categories, the heat map derived from Spearman correlation analysis is less effective and not intuitively clear enough. So here K-means clustering analysis[6] is performed for each vegetable individual product, which is classified into the same category of vegetable individual products with higher correlation[7] .

Firstly, k initial clustering centres $C_i(1 \le i \le k)$ are randomly selected from the dataset, and the Euclidean distance between the remaining data objects and the clustering centres C_i is calculated[8] :

$$d(x, C_i) = \sqrt{\sum_{j}^{m} (x_j - C_{ij})^2} \tag{2}$$

Where, x is the data object, C_i is the i th clustering centre, m is the dimension of the data object, x_j C_{ij} is the j th attribute value of x and C_i .

The sum of squared errors SSE for the entire dataset is calculated as:

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} \left| d(x, C_i) \right|^2 \tag{3}$$

Where the size of SSE indicates the goodness of the clustering result and is the number of clusters.

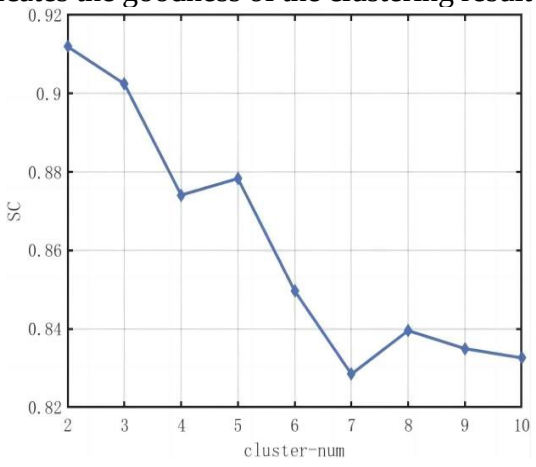


Figure 5: Contour coefficient map

The contour coefficient is an important indicator for evaluating the clustering results, the closer the contour coefficient is to 1, the better the clustering effect is[9]. Although the contour coefficient is closer to 1 when the number of classifications is 2, it lacks practical significance because the number of classifications is too small. Therefore, this paper selects the number of categories when the contour coefficient is 0.902,the results are shown in Fig. 5. i.e., each single product is divided into three categories. The calculated clustering centre of category 1 is 588, the clustering centre of category 2 is 25510, and the clustering centre of category 3 is 8141[10]. The results are shown in Fig. 6.

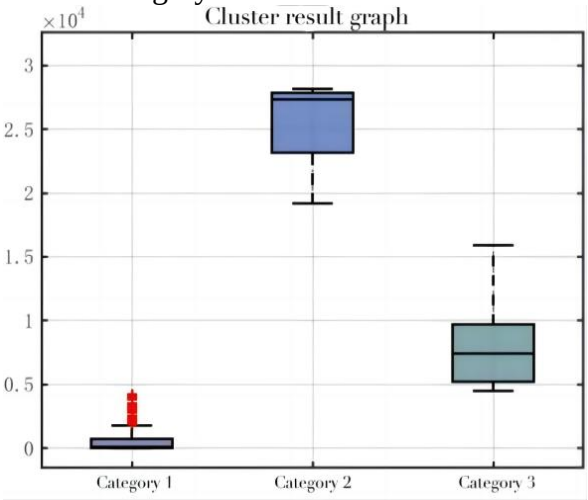


Figure 6: Graph of clustering results

The percentage of categories is then analysed. There were 212 items in category 1, 4 items in category 2 and 30 items in category 3. After analysis it was found that the highest number of clusters were common dishes, the second highest were side dish dishes and the lowest number were other dishes. The percentage is shown in Fig 7, category 1 has the largest percentage, about 86%. Category II has the smallest share, about 2 percent.

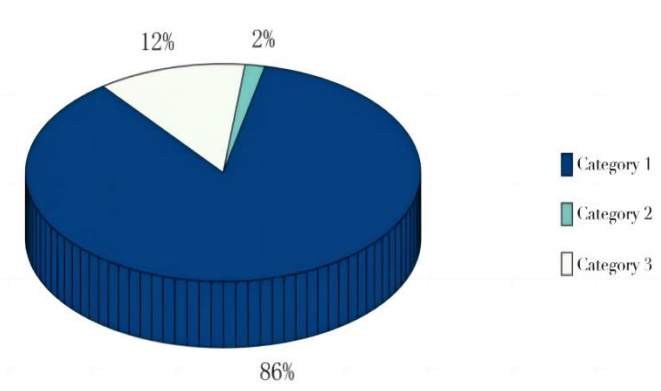


Figure 7: Percentage of each category

6. Conclusions

A large amount of price data of vegetables provide a basis for studying the distribution rules of single products and categories of vegetables. This paper analyzed the time distribution of vegetable sales using time series images and found that the sales tend to reach the maximum in winter, which is related to the inability of vegetables to grow in winter. Spearman The correlation coefficient provides a tool for the study of the correlation between vegetable categories. By studying the correlation between the vegetable categories, the bundling sales of each vegetable category is conducive to

improving the profit of merchants. K-means cluster analysis analyzes the correlation between single vegetable varieties, which facilitates merchants to purchase and sell according to the correlation. The calculation results agree with the actual situation, indicating that the model has broad applicability. In the current period of e-commerce, it is conducive to small and micro businesses to better calculate the amount of food, in order to achieve the purpose of maximizing business profits.

References

[1] Li Zhikun, Jia Wenhua, Zhu Wei, et al. Effect of nitrogen fertilizer and knosteramide on the yield and time distribution of fiber yield and quality in cotton [J]. The Journal of the Crop Sciences.

[2] Pan Haiwei, Hao Longfei, Qin Focang, etc. Seasonal dynamics of soil microbial community structure and metabolic characteristics of pinus tabulaeformis plantation in arsenic sandstone area [J]. Resources and environment in arid areas, 2023, 37(10):168-174.

[3] Chen Jixiang, Yin Haibo, Wang Dan, etc. The relationship between traits and environmental factors of two plants in Liaoning based on gray correlation and Pearson correlation coefficient [J]. Research and Practice of Modern Traditional Chinese Medicine, 2022, 36(01):11-17.

[4] Sun Zhanlong, Zhao Baolong, Bo Yue Xuejing. Application of stability control of coal preparation quality based on dispersion and normal distribution [J]. Coal processing and comprehensive utilization, 2022 (4): 37-41 + 46.

[5] Yun Tao, Zhang Jinnan, Li Shanshan, etc. Exploration of expert review behavior analysis and evaluation method based on Spearman's coefficient and multi-layer perceptron [J]. Science and technology bulletin. 2022, 38(05):107-112.

[6] Wu Huihui, Yuan Zhe, Hui Xiaojian, etc. Analysis of Olympic award-winning studies based on K-means clustering [J]. Modern Information Technology, 2023, 7 (15): 136-140.

[7] Gou Wenjing, Yang Hanjiu, Li Hang. Correlation analysis of the selection of regions of interest and biological prognosis factors for MR diffusion-weighted imaging in breast cancer patients [J]. Journal of Clinical Radiology, 2022, 41(02):251-255.

[8] Xie Yaqi, Miao Yang, Liang Wei, et al. Restoration of shredded paper splicing based on cluster analysis and Euclidean distance model [J]. Electronic Technology and Software Engineering, 2020 (18): 145-146.

[9] Sun Lin, Liu Menghan, Xu Jiucheng. A K-means clustering algorithm [J] based on the optimized initial clustering center and contour coefficients. Fuzzy Systems and Mathematics, 2022, 36(1):47-65.

[10] Luo Shuwen, Wan Renxia, Miao Chuoqian. A three-branch decision density peak clustering algorithm based on the cluster center preselection strategy [J]. Journal of Shanxi University (Natural Science Edition). 1):47-65.