Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 12, No. 1, 2024

Research on Vegetable Pricing and Replenishment Model Based on Greedy Algorithm

Yujing Rao [#], Canbin Wang [#]

School of Statistics and Mathematics, Guangdong University of Finance and Economics, Guangzhou, 510320, China
#These authors contributed equally

Abstract: This article analyzes the pricing and replenishment decisions of supermarkets, exploring the distribution patterns and correlations between different vegetable categories and individual products. On the other hand, a robust linear programming model is constructed, which can iteratively optimize the corresponding replenishment plan, significantly improving the profitability of supermarkets. This provides important practical significance for the operation and sales of vegetable products in fresh food supermarkets. Through Spearman correlation coefficient analysis, it was found that there is a close relationship between cauliflower and leafy vegetables. The correlation between various vegetable items was processed using K-means clustering method, and the correlation between each item was discussed after classifying the groups. Finally, the grey correlation analysis method was used to further explore the ranking of the correlation between each category and individual product.

Keywords: Spearman Correlation Coefficient; K-means Clustering Method; Grey Correlation Analysis Method.

1. Introduction

With the improvement of quality of life, people's requirements for the quality of vegetables are increasing. However, the shelf life of vegetable products is relatively short, and their appearance will gradually deteriorate[1]. For the sales of vegetable products, whether merchants can timely and intact sell vegetable products based on the pricing and replenishment relationship of vegetable products.

In fresh food supermarkets, there are various types and origins of vegetables, and the time for purchasing and trading vegetables is usually in the early morning. Therefore, merchants need to specify different or similar replenishment decisions based on sales situation, people's preferences, historical experience, etc., including the type, origin, quantity, and price of the purchased vegetables. At the same time, sales strategies are formulated based on differences in purchasing locations or prices, and the pricing is generally based on the "cost plus pricing" method. At the same time, vegetables may also experience uncontrollable factors such as transportation damage or poor-quality during transportation, and supermarkets need to conduct statistics on them for discounted sales[2].

Therefore, this article is based on the information of six vegetable categories and multiple vegetable types distributed by a supermarket, the sales situation of vegetable products from July 2020 to June 2023, wholesale data, and recent loss rate data of each product[3]. Based on the actual situation, a mathematical model is established to analyze the automatic pricing and replenishment decisions of supermarket vegetable products and provide suggestions.

2. Distribution Patterns and Correlation Analysis of Various Vegetable Categories

2.1. Distribution Pattern of Vegetable Categories

By classifying the sales volume of six vegetable categories and calculating their total sales volume, descriptive statistical data is obtained as shown in Table 1:

Table 1. Descriptive statistical data for six vegetable categories

Variable Name	Maximum value	minimum value	average value	standard deviation
sales volume	198520.978	22431.782	78495.986	64051.532
Variable Name	median	variance	kurtosis	skewness
sales volume	58926.588	4102598720.177	2.955	1.657

The sales volume and time of vegetable products generally have a certain distribution pattern. By filtering different months, the sales data of different categories and months are obtained, and a line chart is drawn. At the same time, the above data is visualized, as shown in Fig.1:

As shown in fig.1, there is a certain distribution pattern in the sales volume of various categories of vegetables. For example, the maximum sales volume occurs in August, and except for edible mushrooms, the sales volume of other qualities begins to decline after August. In the first two quarters, the sales of the variety showed a general downward trend. The sales trends of edible mushrooms and aquatic rhizomes, as well as the sales trends of flowers, leaves, and chili peppers, are roughly the same.

Filter out sales data of the same variety in different years and draw a bar chart, as shown in Fig.2:

By comparison, it can be concluded that all varieties have reached their highest sales figures in 2022, with the growth rate of sales of flowers, leaves, and chili peppers being relatively fast, leading to an increase in market demand.

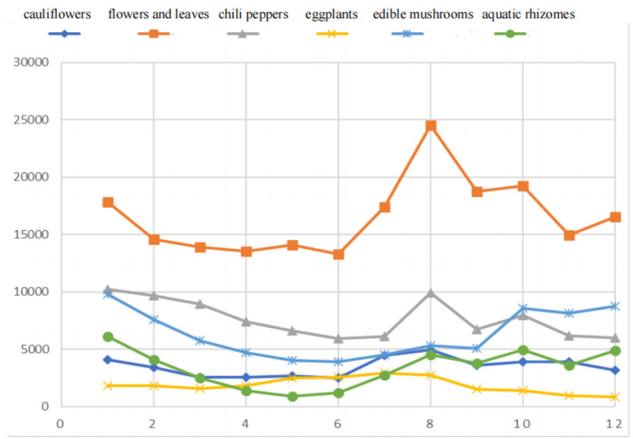

Fig 1. Changes in sales volume of different varieties with months

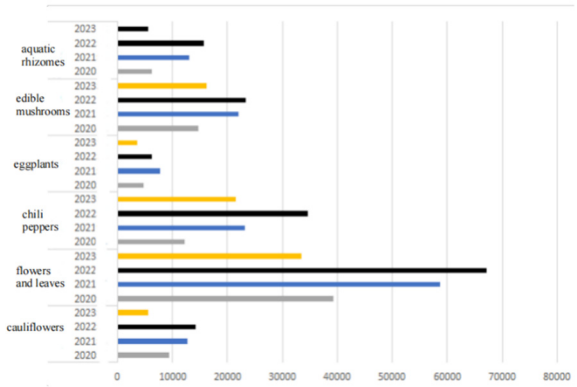

Fig 2. Sales figures of different varieties in different years

2.2. Correlation of Vegetable Categories

Spearman correlation analysis, also known as rank correlation analysis, reflects the correlation between the trend direction and strength of two random variables. It is a statistical measure obtained by arranging the sample values of two random variables in order of data size, and replacing the actual data with the positions of the sample values of each element[4]. The Spearman correlation coefficient is defined as the Pearson correlation coefficient between hierarchical variables.

In the correlation analysis of vegetable categories, different categories of vegetables have different sales volumes and changes in magnitude each month. Therefore, the Spearman correlation analysis method can effectively replace the actual data by substituting the sample values of each category of vegetables, and thus obtain the corresponding correlation coefficients between different categories of vegetables to infer correlations. Definition: and are two sets of data, and the calculation formula for Spearman correlation coefficient is as follows:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \quad (1)$$

Among them, is the rank difference between and (the rank of a number, which is the position of the number in the column sorted from small to large)[5], is the total number of observed samples (there are 36 months of data here), and the correlation coefficient is represented by, which is between -1 and 1.

Therefore, based on Spearman's principle of correlation, a heatmap of the correlation coefficient for the six major vegetable categories of flowers and leaves, cauliflower, aquatic rhizomes, eggplants, chili peppers, and edible mushrooms was created, as shown in Fig.3:

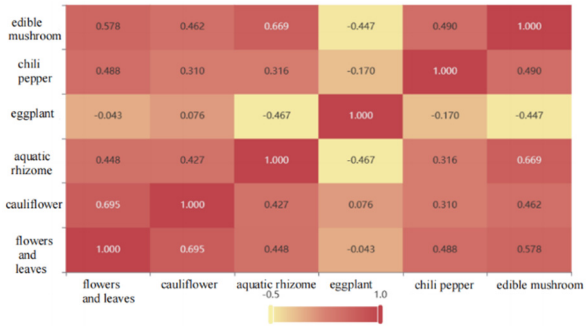

Fig 3. Heat map of correlation coefficients for various vegetable categories

Through a visual presentation of the correlation between various vegetable categories in the heatmap, it is not difficult to determine that on the basis of a positive correlation dimension, the relationship between cauliflower and flower leaf vegetable categories is the closest. Among them, $r_s = 0.695$ when the sales of cauliflower vegetable products are good, the sales of flower leaf vegetable products will also be better.

On the basis of negative correlation dimensions, there is a close relationship between the two vegetable categories of eggplants and aquatic rhizomes, with $r_s = -0.467$indicating that when eggplants sell well, aquatic rhizome products sell generally. Due to the absolute value of r_s being less than 0.5, there is a certain correlation between the two, but the relationship is weak.

In addition, the correlation coefficient between the two vegetable categories of eggplants and cauliflower is $r_s = 0.076$, and the absolute value of r_s is very close to 0, indicating that there is basically no correlation between the two.

3. Distribution Patterns and Correlations of Individual Vegetable Products

By categorizing and summarizing the total sales volume of various vegetable items, corresponding descriptive statistical data can be obtained:

Table 2. Descriptive Statistical Data

average	3139.839	kurtosis	8.465634
Standard error	498.0615	skewness	2.854945
median	356.5045	region	30235.04
Mode	3	minimum value	0.419
standard deviation	6099.982	Maximum value	30235.46
variance	37209783	Summation	470975.9
Observations	150		

The statistics shown in Table 2 can provide an overall understanding of the quantitative data. Between 2020 and 2023, the average total sales volume of all vegetable items was 3139.839kg, with a standard deviation of 6099.982. It can be observed that there is a significant difference in sales volume among various vegetable items.

The total sales volume of the best-selling vegetable item is 30235.46kg, while the total sales volume of the most difficult to sell vegetable item is 0.419kg. After checking the pivot chart, it was found that Yunnan lettuce and red oak leaves were the two, with a significant difference in sales volume. During the period of recorded sales volume from 2020 to 2023, the total sales volume of various vegetable items in the supermarket was 470975.9kg.

3.1. Seasonal Distribution Pattern

Through the data pivot chart in the Excel toolbox, it is not difficult to find that there are significant differences in the sales volume of many vegetable items in certain months after filtering the time. Therefore, it is inferred that the production and sales of vegetables have seasonality, that is, each season has corresponding seasonal dishes[6]. Therefore, we analyzed and studied the seasons of spring, summer, autumn, and winter (1, 2, 3, 4) to find the distribution pattern of the sales volume of vegetable items:

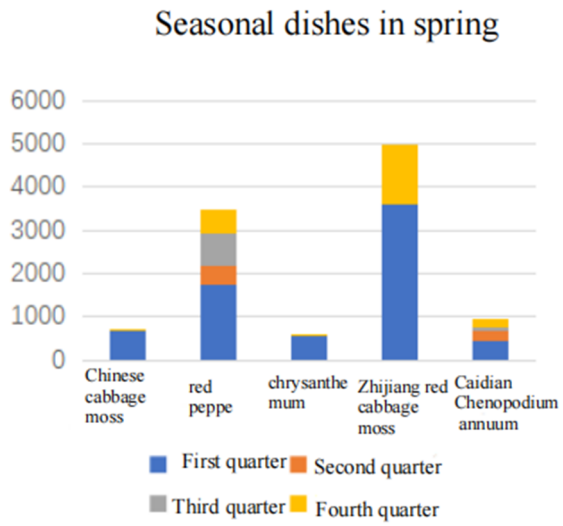

Fig 4. Spring Hot Selling Dishes

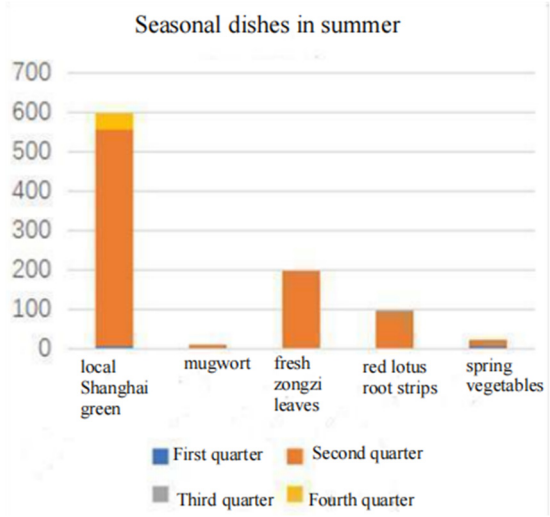

Fig 5. Summer Hot Selling Dishes

From Fig 4 and 5, it can be seen that for example, the sales volume and proportion of five single vegetable products, namely Chinese cabbage moss, red pepper, chrysanthemum, Zhijiang red cabbage moss, and Caidian Chenopodium annuum, in spring are more prominent in all quarters of the year. Taking the local Shanghai green, mugwort, fresh zongzi leaves, red lotus root strips, and spring vegetables as examples, their sales volume in summer accounts for a prominent proportion of the sales volume in all quarters of the year.

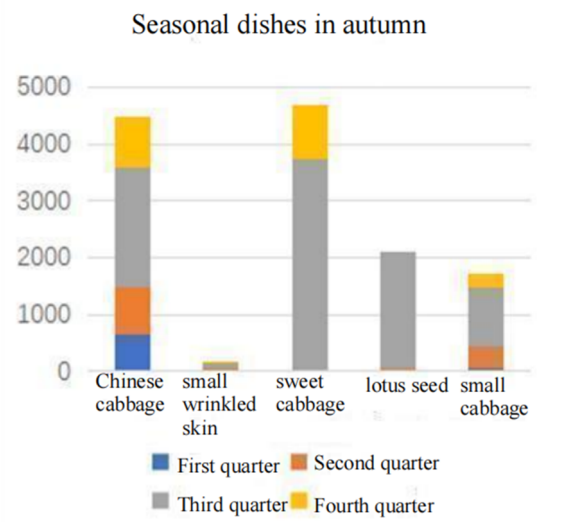

Fig 6. Hot selling autumn dishes

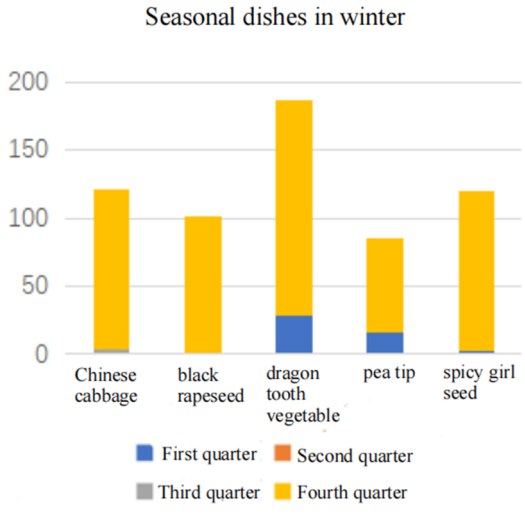

Fig 7. Winter Hot Selling Dishes

From Fig 6 and 7, it can be seen that for example, the sales volume and proportion of the five vegetable items of Chinese cabbage, small wrinkled skin, sweet cabbage, lotus seed, and small cabbage in autumn are relatively prominent in all quarters of the year. Taking five local vegetables, namely Chinese cabbage, black rapeseed, dragon tooth vegetable, pea tip, and spicy girl seed, as an example, their sales volume in winter accounts for a prominent proportion of the sales volume in all quarters of the year.

In summary, due to the influence of seasonality, the sales volume of seasonal vegetables varies greatly in different quarters, usually with a significant distribution in one or two

quarters. However, the planting and production of other vegetable items such as green and red pepper and green stem scattered flowers have a lower dependence on seasons, and the distribution results show that the sales volume difference in each quarter is not significant, so there is different distribution patterns compared to seasonal vegetable items.

3.2. The Correlation between Individual Vegetable Products

The core objective of K-means clustering is to divide a given dataset into K clusters and provide the corresponding center point for each sample data. Its basic idea is to iteratively find a partitioning scheme for K clusters, so as to minimize the loss function corresponding to the clustering results[7]. Among them, the loss function can be defined as the sum of squared errors of the distances between each sample and the center point of the cluster it belongs to:

$$J(c,\mu) = \sum_{i=1}^{M} \left\| x_i - \mu_{C_i} \right\|^2 \quad (2)$$

Among them, x_i represents the th sample, c_i is the cluster to which x_i belongs, μ_{C_i} represents the center point corresponding to the cluster, and M is the total number of samples. The selection of K value is generally based on experimental and multiple experimental results[8]. In this article, there are many types of single products, which is not conducive to analyzing the correlation. Therefore, clustering method is considered to classify single products into different categories. Cluster analysis divides all samples into 4 categories based on data features, and the clustering results are shown in Table 3:

Table 3. Cluster Classification Results Table

category	category1(n=25)	category2(n=5)	category3(n=111)
sales volume	6046.29±1833.80	28427.65±1109.9	445±695.59
category	category4(n=9)	F	P
sales volume	14249.21±3166.48	1182.99	0.000***
Note: * * *, * *, * represent significance levels of 1%, 5%, and 10%, respectively			

From Table 3, it can be seen that for variable sales, the significance P-value is 0.000 * *, which shows significance at the horizontal level. The null hypothesis is rejected, indicating that there is a significant difference in variable sales among the categories classified by cluster analysis. And the contour coefficient is calculated to be 0.792. For a sample set, its contour coefficient is the average of all sample contour coefficients. The range of contour coefficient values is [-1,1]. The closer the distance between samples of the same category, the farther the distance between samples of different categories, the higher the score, and the better the clustering effect[9]. This indicates that the clustering effect of the sample is good. Convert all individual products into four categories as shown in Table 4:

Table 4. Partial Clustering Results Table

vegetable	sales volume	Cluster types
Artemisia argyi	10.512	category3
Cabbage moss	718.676	category3
Artemisia annua	2.224	category3
White Jade Mushroom	3368	category1
Mint leaves	7.568	category3
Baokang Mountain Chinese Cabbage	6484.736	category1
Local yellow heart rapeseed	1375.122	category3
Local Shanghai Qing	596.697	category3
Local Chinese Cabbage	121.02	category3
Golden needle mushroom	28640.79	category2

After dividing each category into four categories, select the five best-selling items in reverse order from the four categories and calculate their standard deviations. Use Pearson correlation coefficient to calculate the relationship between standard deviations of different categories[10]. The Pearson correlation coefficient between two variables is defined as the quotient of the covariance and standard deviation between the two variables:

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \quad (3)$$

The above equation defines the overall correlation coefficient, commonly represented by the Greek lowercase letter ρ. Estimate the covariance and standard deviation of the sample to obtain the Pearson correlation coefficient[11], commonly represented by the lowercase letter r in English:

$$r = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}} \quad (4)$$

The varieties selected in category 1 are seafood mushrooms, bubble peppers (premium), green stem scattered flowers, baby bok choy, and bamboo leaf vegetables. The correlation coefficient heatmap is shown in Fig.8:

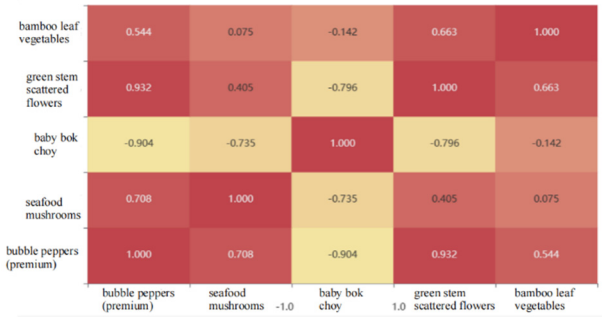

Fig 8. Thermal diagram of correlation coefficient for category 1

The varieties selected in category 2 are golden needle mushroom, clean lotus root, Wuhu green pepper, broccoli, and Yunnan lettuce. The correlation coefficient heatmap is shown in Fig.9:

The varieties selected in category 3 are Pleurotus ostreatus, Xixia shiitake mushroom, Hibiscus erinaceus, Lentinula edodes, and Cordyceps sinensis. The correlation coefficient heatmap is shown in Fig.10:

The varieties selected in category 4 are spinach, Chinese cabbage, Sichuan pepper, milk cabbage, and Shanghai green. The correlation coefficient heatmap is shown in Fig.11:

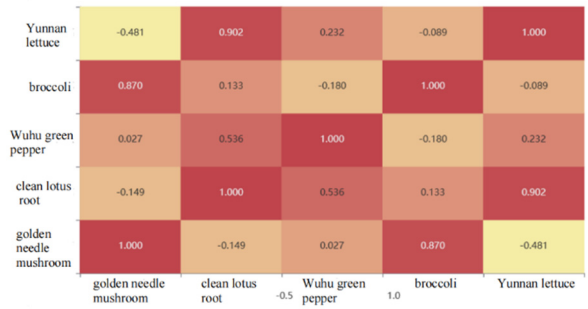

Fig 9. Thermal diagram of correlation coefficient for category 2

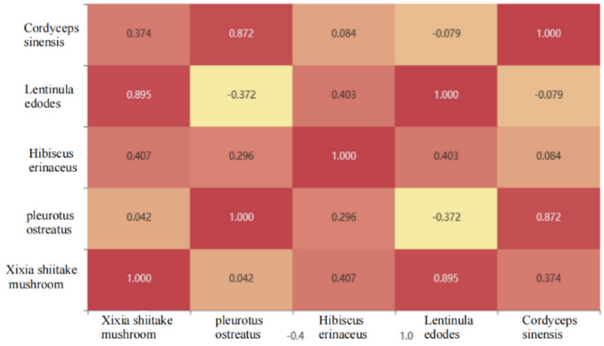

Fig 10. Heat map of correlation coefficient for category 3

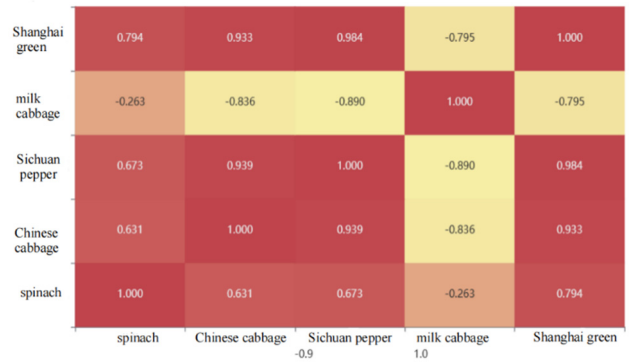

Fig 11. Heat map of correlation coefficient for category 4

The higher the correlation coefficient, the greater the correlation between the standard deviations of the two individual products, that is, the more similar the annual fluctuation level, and the two show a positive correlation change. By analyzing the heatmaps of various correlation coefficients, it can be concluded that there is a strong correlation between pickled pepper and green stem scattered flowers, clean lotus root and Yunnan lettuce, long line eggplant and Xixia mushroom, screw pepper and Chinese cabbage, Chinese cabbage and Shanghai green, and screw pepper and Shanghai green. That is, the two varieties of vegetables can be purchased well at the same time, forming a positive correlation. There is a negative correlation between baby bok choy and pickled peppers, as well as between milk cabbage and screw peppers. When one vegetable has high sales, the other will inevitably have poor sales.

4. Conclusion

This article uses the dataset obtained from data preprocessing to provide descriptive statistical explanations for six major categories of vegetable products and various types of vegetable single products. Histograms are drawn to visualize the time distribution patterns of each category, and bar charts are drawn to discover seasonal differences between vegetable single products. Through Spearman correlation coefficient analysis, it was found that there is a close relationship between cauliflower and leafy vegetables. At the same time, the K-means clustering analysis method is used to first distinguish the categories of various vegetable items, reducing the sample size of supermarket vegetable items and avoiding the cumbersome correlation analysis between items due to the large sample size. It can also enhance the visibility of correlations between projects. By using nonlinear regression equations to fit the relationship between total sales and cost plus pricing for each vegetable category, the problem of poor fitting accuracy caused by scattered data samples can be overcome.

References

[1] Jiang Xinyi, Zhou Hui, Hu Xiangdong. Current situation, problems and countermeasures of vegetable circulation system in China [J] . Agricultural outlook, 2024,20(04) : 3-10.

[2] Liu Hengyu. Study on the strategy of fruit and vegetable product inventory, production and sale [D] . Beijing Jiaotong University, 2019.

[3] Eugene. A study on joint replenishment and cost allocation strategy for perishable products with time-varying deterioration rate [D] . Southeast University, 2024.

[4] company. Spearman rank correlation coefficient of integrated interval number and its application [J] . Journal of Chongqing Technology and Business University science, 2020,37(06) : 71-75

[5] Qian Chenjian. Hypothesis test for comparison of Spearman rank correlation coefficients and a new method for sample size estimation [D] . Southern Medical University, 2022.

[6] Yang Juan, Qian Tingting, Zheng Xiuguo, etc. Characteristics and influencing factors of national and regional vegetable price trend [J]. Journal of China Agricultural University, 2021,26(02) : 188-198.

[7] Liu Jianhua. Improvement and application of K-means clustering algorithm [J] . Journal of Taiyuan Normal University Science, 2020,19(01) : 81-83.

[8] Yang keyi, Bao liangqi, Zhao Jun, etc.Improved model compression algorithm based on K-means clustering [J] . Information technology and informatization, 2022(04) : 212-215.

[9] Sun Lin, Liu Menghan, Xu Jiucheng. K-means clustering algorithm based on optimizing initial clustering center and contour coefficient [J]. Fuzzy systems and mathematics, 2022,36(01) : 47-65.

[10] Zhao Yuanshang, Lin Weifang. Typical scenarios based on Pearson product-moment correlation coefficient fusion density peaks and entropy weight method [J] . China electric power, 2023,56(05) : 193-202.

[11] Cheng Kuen-kuen. An empirical study on the relationship between scientific research and teaching in universities -- an analysis based on Pearson product-moment correlation coefficient [J] . Science and Technology in Chinese universities, 2022(10) : 46 -52.