Research on the correlation and distribution law of commodities

Dingfeng Zhang[1, #, *], Yuhang She[2, #], Limei Liu[2, #], Rongjie Ouyang[1, #], Jinpeng Ye[3, #]

[1]College of Engineering, Shantou University, Shantou, China, 515063

[2]College of Business, Shantou University, Shantou, China, 515821

[3]Department of college of mathematics and computer, Shantou University, Shantou, China, 515821

* Corresponding Author Email: 21dfzhang1@stu.edu.cn

#These authors contributed equally.

Abstract. In the fresh supermarket, the supermarket will generally make replenishment plans according to the historical sales and demand of each commodity. There is a certain correlation between the sales volume of different commodities. It is crucial to deeply understand the distribution law and correlation relationship of commodity sales to optimize the replenishment strategy. In this paper, Spearman correlation coefficient analysis and entropy weight method-TOPSIS model are constructed to analyze the distribution law and correlation relationship of vegetable products and single product sales volume. First of all, this paper aggregates the sales volume of different vegetable categories and single items by month to get the average monthly sales volume. The average monthly sales data of each vegetable category was visualized, and the distribution pattern of each vegetable category was analyzed. The relation of vegetable categories was analyzed using Spearman's correlation coefficient. Then, for vegetable items, the TOPSIS model based on entropy weight method was constructed, and the 8 most representative items were found among 251 vegetable items. Finally, the average monthly sales data of the selected vegetable products was visualized, the distribution pattern of the sales of each vegetable product was analyzed and the correlation coefficient of Spearman was used to analyze the sales of different vegetables.

Keywords: Spearman Correlation Coefficient, Entropy Weight Method Modified TOPSIS Evaluation Method, Data Visualization.

1. Introduction

Replenishment of commodities is the key content of commercial super to achieve operating profit[1]. Either too much or too little replenishment of commodities is not conducive to the maximization of business benefits for supermarkets. In order to formulate appropriate pricing and replenishment strategies, replenishment reference can be made through the correlation of distribution laws of commodity sales, that is, the correlation of pair commodity sales can be analyzed. For the correlation analysis of commodities, R. grawal et al first used Apriori to conduct correlation analysis of commodities in 1993 [2]. In China, Chen Danchun et al. used grey relational degree to analyze the relevant factors affecting the sales volume of commodities [3]. Xue Liang et al. used Pearson correlation coefficient to calculate the correlation of products selected by different users[4]. Zhang Mengni et al. used principal component analysis to conduct comprehensive evaluation of commodities [5], which can be applied to the selection of typical commodities.

However, Apriori algorithm has low efficiency[6] and narrow adaptation range. Grey relational degree analysis requires the data to be grey system, which is not suitable for popularization. Pearson correlation coefficient requires the data to be normal and ppairs present a linear relationship [7], which has a small scope of application. After principal component analysis, the explanation of new variables is poor, and TOPSIS can objectively evaluate the scoring grade of specific objects [8].

Therefore, this paper adopts the widely applicable Spearman correlation coefficient to conduct commodity correlation analysis, and analyzes the distribution and correlation of the sales volume of various vegetable categories and vegetable items. Due to the large quantity of vegetable items, this paper adopts TOPSIS comprehensive evaluation method modified by entropy weight method to screen out 8 typical vegetable items, and then analyzes the distribution law and correlation relationship of vegetable item sales.

(Data source:http://www.mcm.edu.cn/html_cn/node/c74d72127066f510a5723a94b5323a26.html)

2. Distribution pattern and correlation analysis of sales volume of vegetable categories

2.1. Data preprocessing: vegetable category sales volume aggregated by month

The data includes the sales details of each commodity and the relevant data of wholesale price of the supermarket from July 1, 2020 to June 30, 2023[9], and the data is huge and messy, which is not conducive to analysis. Therefore, this paper uses Excel tool to process the original data and aggregates the sales data of each vegetable category on a monthly basis to make data analysis more efficient. In order to analyze the distribution pattern of vegetable sales by month and month, this paper averages the sales data of all kinds of vegetables from 2020-20-2023 by month, and obtains the average sales volume of all kinds of vegetables in different months in these three years as shown in Table 1.

Table 1. Average annual sales of six vegetable categories

	florifolias class	aquatic rhizome class	eggplant class	chilli class	edible fungi class	cauliflower class
Jan	5951.489	2030.08	599.5653	3399.807	3274.821	1367.771
Feb	4854.31	1353.578	616.7483	3227.877	2530.217	1145.678
Mar	4632.147	838.2573	518.918	2968.657	1909.341	859.191
Apr	4515.549	464.4687	609.566	2463.441	1561.404	844.248
May	4701.009	295.5937	825.8353	2205.674	1351.573	901.638
Jun	4431.186	399.4167	855.9317	1980.076	1303.518	831.6697
Jul	5790.769	913.4457	966.9737	2045.61	1504.409	1489.632
Aug	8168.644	1502.366	912.063	3295.511	1765.584	1642.145
Sep	6241.422	1256.564	505.8567	2251.268	1685.908	1196.13
Oct	6404.062	1644.864	466.877	2644.513	2846.258	1301.774
Nov	4977.301	1195.65	312.1053	2053.31	2718.937	1293.832
Dec	5505.771	1632.833	286.8203	1993.799	2910.272	1048.441

The above obtained the monthly sales volume data of each category of vegetables and the average monthly sales volume data obtained by taking the mean value by month from 2020 to 2023, and utilized the obtained monthly average sales volume data to carry out the distribution law of the sales volume distribution of different categories of vegetables and the analysis of correlation.

2.2. Modeling

2.2.1 Analysis of the distribution law of sales volume of vegetable categories

This paper uses the average monthly sales volume data of each vegetable category obtained from processing to construct the curve diagram of its average monthly sales volume change with the month, and explores the distribution law of sales volume of different vegetable categories by adopting the means of data visualization and combining with the analysis of data characteristics.

2.2.2 Spearman correlation coefficient

In this paper, the Spearman rank correlation coefficient was used to measure the correlation between different vegetable categories. Spearman Rank correlation coefficient is a non-parametric rank correlation analysis method used to analyze the correlation between two variables. The Spearman rank correlation coefficient is calculated as [10]:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{N(N^2-1)} \tag{1}$$

Where: ρ denotes the Spearman's correlation coefficient between the two vegetable categories, $d_i = X_i - Y_i (1 \leq i \leq N)$, denotes the difference between the rank values of the ith data pair, and N denotes the total number of observed samples.

2.2.3 Correlation coefficient matrix

Through Spearman's correlation analysis, this paper calculates the correlation coefficients between the average annual sales of six vegetable categories, constituting a 6×6 matrix of correlation coefficients P. This matrix clearly describes the degree of correlation between the sales of different categories of vegetables.

$$P = \begin{bmatrix} 1 & \rho_{(1,2)} & \cdots & \rho_{(1,6)} \\ \rho_{(2,1)} & 1 & \cdots & \rho_{(2,6)} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{(2,1)} & \rho_{(6,2)} & \cdots & 1 \end{bmatrix} \tag{2}$$

Equation (2) represents the relationship matrix for the vegetable category. Where $\rho_{(i,j)}$ denotes the correlation coefficient of vegetable category i and vegetable category j.

2.3. Solution of the model

2.3.1 Analysis of the distribution pattern of sales volume of different categories of vegetables

Take the month as the horizontal axis and the sales volume as the vertical axis. The curve diagram of the change of vegetable sales over time is drawn as shown in Fig 1.

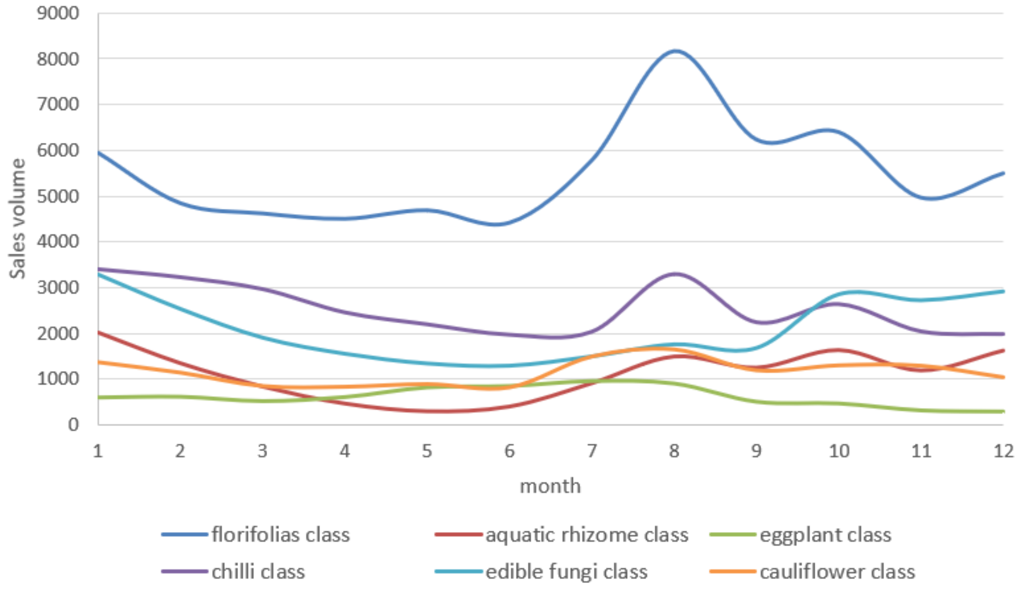

Fig. 1 Trend of sales volume of six major vegetable categories

The distribution pattern of sales volume of different vegetable categories is shown in Fig1. Overall, the sales volume of florifolias class vegetables is much higher than that of other vegetable categories. The sales volume of most vegetable categories had a clear seasonal pattern, with the sales volume of the florifolias class in the second half of the year being significantly higher than that in the first half of the year, which was reflected in the higher sales volume in January, August, September, and October, and the sales volume of the chilli class being at a low point in June, July, and December. aquatic rhizome class vegetables as a whole sold relatively less in the summer months, and the sales trends for both were generally the same, suggesting a strong correlation between them. eggplant class and cauliflower class vegetables both sold less and varied steadily, with less volatility.

2.3.2 Spearman correlation analysis model solution

Using SPSSPRO, Spearman correlation coefficients between different vegetable categories were calculated to obtain the correlation coefficient matrix, which was visualized as a heat map as shown in Fig 2. In this heat map, red indicates positive correlation, blue indicates negative correlation, and larger values and darker colors indicate stronger correlation.

	florifolias class	aquatic rhizome class	eggplant class	chilli class	edible fungi class	cauliflower class
florifolias class	1	0.769	-0.126	0.385	0.462	0.888
aquatic rhizome class	0.769	1	-0.413	0.483	0.86	0.657
eggplant class	-0.126	-0.413	1	0.063	-0.692	0.112
chilli class	0.385	0.483	0.063	1	0.434	0.392
edible fungi class	0.462	0.86	-0.692	0.434	1	0.392
cauliflower class	0.888	0.657	0.112	0.392	0.392	1

Fig. 2 Correlation coefficient relationship graph

The correlation between the sales volume of each vegetable category is obtained from the thermal chart of the correlation coefficient in Fig 2. There is a strong positive correlation between florifolias class and aquatic rhizome class. There is a strong positive correlation between florifolias class and cauliflower class. The correlation between aquatic rhizome class and edible fungi class is very strong. There is a certain negative correlation between eggplant class and edible fungi class. The correlation between eggplant class and other categories was weak, and the correlation between eggplant class and chilli class was close to 0.

3. Analysis of distribution rule and correlation of vegetable sales

3.1. Analysis of distribution rules

The business super sales a total of six vegetables category, each category and covers a large number of items, in order to reduce the data dimension, this paper chose based on entropy correction of TOPSIS evaluation [11], for 251 vegetables item each item score, and then select the highest score of eight vegetable item for further correlation analysis.

3.1.1 Modified TOPSIS model based on entropy weight method

Similar to processing the sales data of the six vegetable categories, this paper aggregates the sales data of 251 vegetable individual categories by month, aggregating them into January-December data. Then the data for the three years from 2020-2023 are averaged to get the average monthly sales for each month. Therefore, the indicators to be evaluated are the sales data of the vegetable single category from January to December, and there are 12 indicators in total.

There are n objects to be evaluated and m evaluation indicators form the normalization matrix as:

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \tag{3}$$

Where n=251,m=12,the sales data of vegetables are $x_{ij}(i=1,2,\ldots,n;j=1,2,\ldots,12)$

First of all, the indicator data should be normalized and normalized to calculate the normalization matrix:

$$r_{ij}=\frac{x_{ij}-x_j^{min}}{x_j^{max}-x_j^{min}} \tag{4}$$

where $x_j^{max}=\max\{x_{i1},x_{i2},\ldots,x_{im}\}$, $x_j^{min}=\min\{x_{i1},x_{i2},\ldots,x_{im}\}$, the normalized matrix is denoted as $R=[r_{ij}]_{n\times m}$

Calculate the weight of the i^{th} sample under the j^{th} indicator and regard it as the probability used in the relative entropy calculation, the formula of the probability matrix is:

$$p_{ij}=\frac{r_{ij}}{\sum_{i=1}^{n} r_{ij}} \tag{5}$$

The information entropy of each indicator is calculated and the information utility value is calculated and normalized to get the entropy weight of each indicator. The formula for information entropy is:

$$E_j=-\frac{1}{\ln m}\sum_{i=1}^{m} p_{ij}\ln p_{ij} \tag{6}$$

Utilizing the information entropy in equation (6), the weight of the indicator is obtained as follows.

$$\omega_j=\frac{1-E_j}{\sum_{j=1}^{n}(1-E_j)} \tag{7}$$

TOPSIS model is a comprehensive evaluation method that ranks evaluation objects by comparing their proximity to the idealized target [12]. In this paper, TOPSIS model is used to select the most representative vegetable single product.

Firstly, the positive ideal solution of each index, i.e., the maximum value of the sales volume of vegetable single products in each month, is found out, which is denoted as$z_i^+(i=1,2,\ldots,m)$, and composes the vector

$$Z^+=\{z_1^+,z_2^+,\ldots,z_m^+\} \tag{8}$$

Find the negative ideal solution of each indicator, i.e., the lowest value of vegetable unit sales in each month, denoted as $z_i^-(i=1,2,\ldots,m)$, form the vector

$$Z^-=\{z_1^-,z_2^-,\ldots,z_m^-\} \tag{9}$$

Then the distance of each sample from the positive ideal solution is calculated, where D_i^+ denotes the distance of the ith vegetable item from the ideal optimal objective.

$$D_i^+=\sqrt{\sum_{j=1}^{m}\left(z_{ij}-z_j^+\right)^2} \tag{10}$$

Calculate the distance of each sample from the negative ideal solution, where D_i^- denotes the distance of i vegetable singles from the ideal worst goal.

$$D_i^- = \sqrt{\sum_{j=1}^{m} \left(z_{ij} - z_j^-\right)^2} \tag{11}$$

Finally, the score of the ith vegetable item is calculated:

$$C_i = \frac{D_i^-}{D_i^+ + D_i^-} \tag{12}$$

Where the value of C_i ranges from [0,1], the closer C_i is to 1, the more representative this vegetable item is. Conversely, the closer the value of C_i is to 0, the less representative this vegetable item is.

3.1.2 Analysis of single product sales volume distribution law

Through the entropy weight method modified TOPSIS model to get the score of each vegetable category, this paper screened the highest score of the eight single product, using Excel to build the average monthly sales of vegetable single product with the month of the curve, combined with the characteristics of the data to explore the distribution of the sales volume of different vegetable categories.

3.1.3 Correlation analysis using Spearman correlation coefficient

Firstly, the score of each vegetable item is calculated by the TOPSIS method modified by entropy weight method and sorted from high to low, and the most representative 8 vegetable items are obtained, and the sales data of these 8 vegetable items are used to carry out the Spearman correlation coefficient analysis to find out the correlation coefficient matrix, and then the correlation coefficient matrix is visualized as a heat map, which clearly demonstrates the correlation between different individual items.

3.2. Model Solving

3.2.1 Score calculated based on TOPSIS evaluation method modified by entropy weight method

Using SPSSPRO, the TOPSIS evaluation model modified based on entropy weight method is solved, the score of each vegetable item is calculated and ranked from high to low, and the eight vegetable items with the highest scores are selected as shown in Table 2:

Table 2. Scores of the top 8 vegetable individual items

Score ranking	Vegetable items	Score
1	purple eggplant(2)	0.9845
2	pakchoi	0.9465
3	broccoli	0.9387
4	chinese cabbage	0.8774
5	wild rice stem	0.8497
6	green peppers	0.8310
7	xixia letinous edodes(1)	0.8047
8	enoki mushrooms	0.8014

The 8 vegetable items with the highest scores selected above are the most representative among the 251 vegetable items, so this paper mainly focuses on analyzing the distribution pattern and interrelationship of the sales volume of these 8 vegetable items.

According to the 8 vegetable items obtained in Table 4, similar to the method of dealing with vegetable category data, the data will be aggregated and averaged on a monthly basis to obtain the average annual sales volume of 8 vegetable categories as shown in Table 3:

Table 3. Average annual sales of 8 vegetable categories

	purple eggplant(2)	pakchoi	broccoli	chinese cabbage	wild rice stem	green peppers	xixia letinous edodes(1)	enoki mushrooms
Jan	460.709	1.144667	972.915	157.117	0	47.69833	623.2237	133.4083
Feb	474.497	4.567	794.444	244.1557	2.182667	36.91467	179.0627	153.595
Mar	404.0107	12.60633	510.136	243.34	28.36633	27.47433	184.628	76.46067
Apr	436.2233	9.346333	535.6927	227.1353	44.434	24.838	103.9853	82.95367
May	585.1883	57.96567	639.5403	202.6327	47.12233	21.775	97.88367	62.28433
Jun	477.5287	65.12633	625.387	173.6753	98.70667	5.770667	142.444	55.45933
Jul	445.6723	137.7153	969.7723	306.0337	84.993	52.19067	337.7897	227.3833
Aug	353.7663	116.7857	1008.498	411.1913	44.33433	61.304	317.994	197.1807
Sep	224.6737	89.43767	761.0283	245.8673	24.18367	36.49967	272.1423	163.471
Oct	237.8947	51.099	818.7597	155.026	15.816	33.42533	553.715	169.0633
Nov	212.6133	17.08833	885.1447	88.08833	1.522667	21.762	522.3747	135.9653
Dec	221.223	6.218667	657.758	81.32267	1.555667	25.85767	638.166	108.5983

The above obtained 8 vegetable single product sales volume by monthly average obtained monthly average sales volume, will use the obtained monthly average sales volume data for vegetable single product sales volume distribution pattern and correlation analysis.

3.2.2 Distribution pattern of different single-item vegetables

After evaluation through the TOPSIS evaluation method, the eight highest-scoring vegetable single-item sales were selected, and these eight vegetable single-items contained the vast majority of the information of the 251 single-item data, and the curve graph was drawn as shown in Fig 3.

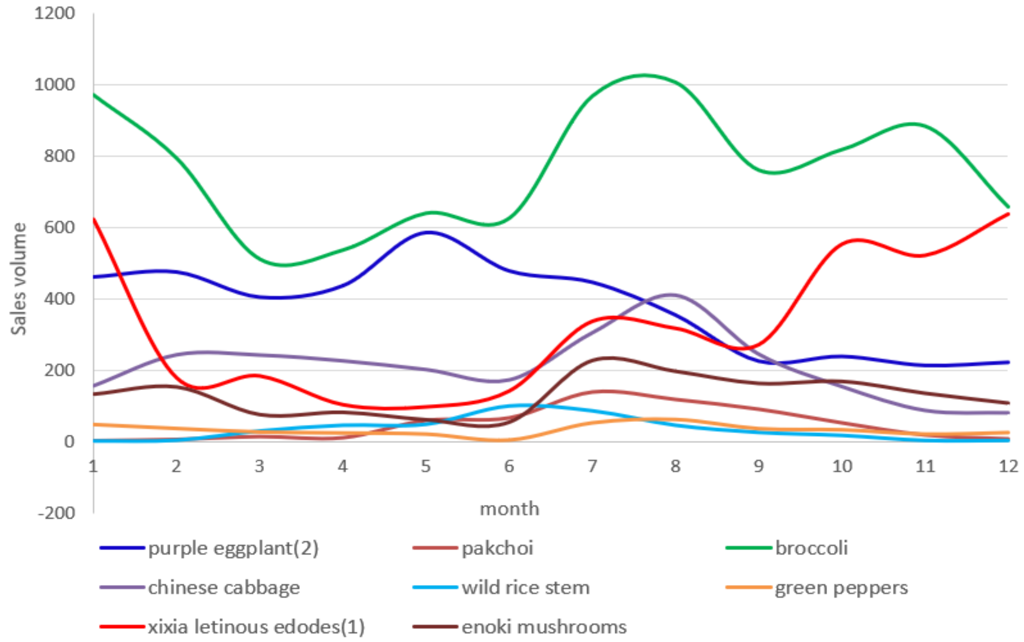

Fig. 3 Trend of sales of 8 individual vegetable items

Overall, there is also a clear seasonal trend in the sales of vegetable items, with peaks in July and August, especially in the sales of "broccoli", "chinese cabbage" and "enoki mushrooms". "enoki mushrooms". For the vegetable items except "purple eggplant (2)", March and April are the peak months with low sales volume. The supply of vegetables at this time is low, and the market share is small. On the whole, the sales volume of vegetables in the first half of the year was less than that in the second half of the year.

There are large differences in sales volume between different vegetable categories. For example, wild rice stem", "pakchoi" and "green peppers" had low and relatively stable sales volumes, while "xixia letinous edodes (1)" and "purple eggplant (2)" had low and relatively stable sales volumes. "purple eggplant (2)" fluctuated considerably.

3.2.3 Spearman's correlation coefficient solution

Using SPSSPRO software, the Spearman correlation coefficient between different vegetable items was solved by two-by-two calculations, the correlation coefficient matrix was solved, and the heat map of correlation coefficient was drawn as shown in Fig 4.

	purple eggplant(2)	pakchoi	broccoli	chinese cabbage	wild rice stem	green peppers	xixia letinous edodes(1)	enoki mushrooms
purple eggplant(2)	1	0. 118	−0. 318	0. 3	0. 755	−0. 082	−0. 791	−0. 373
pakchoi	0. 118	1	0. 455	0. 509	0. 582	0. 291	0. 036	0. 4
broccoli	−0. 318	0. 455	1	0. 3	−0. 236	0. 582	0. 573	0. 855
chinese cabbage	0. 3	0. 509	0. 3	1	0. 409	0. 764	−0. 273	0. 5
wild rice stem	0. 755	0. 582	−0. 236	0. 409	1	−0. 055	−0. 6	−0. 218
green peppers	−0. 082	0. 291	0. 582	0. 764	−0. 055	1	0. 291	0. 836
xixia letinous edodes(1)	−0. 791	0. 036	0. 573	−0. 273	−0. 6	0. 291	1	0. 591
enoki mushrooms	−0. 373	0. 4	0. 855	0. 5	−0. 218	0. 836	0. 591	1

Fig. 4 Plot of correlation coefficient relationship

From the heat map in Fig 4, the correlation coefficient between "wild rice stem" and "purple eggplant (2)" is 0.755, which is a strong correlation, and the correlation coefficient between "enoki mushrooms" and "green peppers" is 0.836, which is a strong correlation." The correlation coefficient between "pakchoi" and "enoki mushrooms" is 0.4, indicating a positive correlation between them. The correlation coefficient between "broccoli" and "chinese cabbage" is 0.3, indicating a positive correlation. The correlation coefficient between "purple eggplant (2)" and "pakchoi" is 0.118, which means that the correlation is weak." The correlation coefficient between "purple eggplant (2)" and "chinese cabbage" is 0.3, which is also weak." The correlation coefficient between "purple eggplant (2)" and "xixia letinous edodes(1)" was -0.791, indicating a strong negative correlation between them.

4. Conclusions

In this paper, by constructing Spearman correlation coefficient and TOPSIS model based on entropy weight method, we analyze the distribution law and correlation relationship of vegetable products and single product sales in fresh suppliers. For vegetable categories, the seasonal characteristics of vegetable category sales were obtained through visual analysis, especially the outstanding performance of florifolias class and chilli class. At the same time, the Spearman correlation coefficient is used to reveal the correlation between different categories, which provides a more comprehensive sales strategy guidance for the supermarket. For vegetable products, 8 most representative vegetable products were selected through TOPSIS model, and the seasonal law of the sales volume of single products and the sales volume of different items were revealed through visualization and Spearman correlation coefficient analysis. The model is objective and true with strong reliability. These research results provide a substantial reference for the replenishment strategy and inventory management of supermarkets, and help them to deal with market changes more flexibly.

References

[1] Haiyun He. A modeling study of back stocking and shelf replenishment of perishables under perceived freshness[D]. Chongqing Jiaotong University,2023.

[2] R. Agrawal, T. Imielinski and A. Swami. Mining association rules between sets ofitems in large databases. Proceedings of the ACM SIGMOD Conference on Management of data,1993:207-216.

[3] Chen Dancun, Liang Meili. Application of BP neural network in used car sales forecasting in China[J]. Statistics and ·Application, 2022, 11: 1029.

[4] Xue Liang, Xu Hui, Feng Zunlei et al. An improved collaborative filtering method for product recommendation[J]. Computer Technology and Development,2022,32(07):201-207.

[5] ZHANG Mengni, CHU Yuping, DONG Xiaodong. Research on supermarket commodity structure based on principal component analysis[J]. Modernization of Shopping Malls,2007(12):44-45.

[6] Tian Jianyong. An Analysis of Association Rule Algorithms in Data Mining--Taking Apriori Algorithm as an Example[J]. Information Technology and Informatization,2020(07):34-35.

[7] LI Yuan, LIU Yutian, FENG Liwei. Nonlinear dynamic process feature extraction and fault detection based on Spearman correlation analysis[J]. Journal of Shandong University of Science and Technology(Natural Science Edition),2023,42(02):98-107.

[8] Shu-Juan Guo, Yadong Xu, Jin-Yong Huang. Evaluation of agricultural green development level based on entropy-weighted TOPSIS model--Taking Henan Province as an example[J/OL]. Journal of Zhejiang University Journal of Jiang University (Agriculture and Life Sciences Edition): 1-10[2023-11-18].

[9] Data source: 2023 National Student Mathematical Modeling Competition official website.

[10] Eden S K, Li C, Shepherd B E. Nonparametric estimation of Spearman's rank correlation with bivariate survival data[J]. Biometrics, 2022, 78(2): 421-434.

[11] Chakraborty S. TOPSIS and Modified TOPSIS: A comparative analysis[J]. Decision Analytics Journal, 2022, 2: 100021.

[12] Pandey V, Komal, Dincer H. A review on TOPSIS method and its extensions for different applications with recent development[J]. Soft Computing, 2023, 27(23): 18011-18039.