Information Systems and Economics (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/infse.2024.050208
ISSN 2523-6407 Vol. 5 Num. 2

A simplification strategy for vegetable supply and marketing type problems based on correlation analysis

Jingyu Qian[1,a], Guangqiang Li[1,b], Shuming Li[1,c], Wenchao Ma[2,d], Yunsheng Zhang[1,e]

[1]Institute of Energy and Mining Engineering, Shandong University of Science and Technology, Qingdao, China

[2]School of Business Administration, Qingdao City University, Qingdao, China

[a]15610589872@163.com, [b]alipang1129@163.com, [c]19863703216@163.com, [d]mawenchao612@163.com, [e]zys15506680208@163.com

Keywords: Statistical analysis, Spearman correlation coefficient, Systematic clustering

Abstract: There is a wide variety of vegetables and the connections between single product sales are complex. Solving the problems related to vegetable supply and sales is difficult. Therefore, it is necessary to simplify the handling of vegetable supply and sales issues to bring greater economic value. The sales of vegetables between the complexity of the connection due to the wide range of vegetables. To solve the problem of vegetable supply and marketing category, we should simplify the processing of vegetable supply and marketing category problems to bring greater economic value. In this paper, the distribution of six vegetable categories of commodity information and sales flow details of the receipt based on a supermarket in the past three years. In addition, the vegetable supply and sales of class problems are analyzed to simplify the strategy. Firstly, eliminating products with minimal sales over the past three years and utilizing box plots to filter out and remove outliers, then filling missing values using spline interpolation, completing the data preprocessing. Secondly, descriptive statistics are introduced for overall analysis to derive and visualize the distribution patterns of various classes. For the category Spearman correlation coefficient matrix, the sales volume is derived and the correlation coefficient matrix is obtained using MATLAB. For the single product correlation coefficient in the data, the systematic clustering is used to simplify it into 8 categories of daily vegetables and seasonal products, and the sales volume of each period is obtained through data analysis and the Spearman correlation coefficient matrix is derived with MATLAB. This paper combines commodity information and sales details, comprehensively interferes with data screening, statistical analysis, correlation analysis to provide a set of strategies simplifying the problem of vegetable supply and marketing. We lay the foundation for the establishment of vegetable optimization class model.

1. Introduction

In daily life, there are various kinds of vegetable commodities [1], and there are also many correlations between different categories and individual products, which makes it very difficult to solve the vegetable sales class problem. Simplifying the vegetable sales class problem lays a certain foundation for the establishment of vegetable-related models, which can better solve the problems of vegetable supply and marketing and pricing, and reduce the economic loss situation caused by vegetables exceeding the freshness period [2]. Scholars at home and abroad have conducted a lot of research on clustering problem to find a simple and practical model. Somayeh Danesh Asgari et al [3] determined the importance of each data point in estimating the trade-off parameter based on the Position-regulated support vector clustering PRSVC to compensate for the information loss and suboptimal solution. Shen [4] proposed a new clustering filtering model based on entropy criterion. Rodriguez [5] used the concept and demonstration of wellhead protection zone delineation in transient aquifers as an example for computationally efficient and goal-oriented uncertainty quantification based on optimized random field clustering. Zhou [6] used a new technique based on rough fuzzy clustering of multi-granularity approximation regions to handle and fuzzily the uncertainty associated with the parameter m.

The above mentioned various optimization and processing of the clustering model to better adapt to the real situation, but none of them applies clustering to simplify the problem. This paper combines commodity information and sales details, and comprehensively interferes with data screening, statistical analysis, correlation analysis and other methods to provide a set of strategies to simplify the problem of vegetable supply and sales. Finally, we lay the foundation for the establishment of vegetable optimization class model.

2. Experimental analysis

Considering the complexity and contingency of the actual situation, a study of the past three years' vegetable sales data by category was conducted based on three basic assumptions, and the logical relationship of the data was statistically analyzed. The assumptions are as follows: (1) It is assumed that vegetables not sold on the same day are not sold on the next day. (2) It is assumed that the selling price of vegetables on holidays is the same as normal. (3) It is assumed that supermarkets maintain stable marketing.

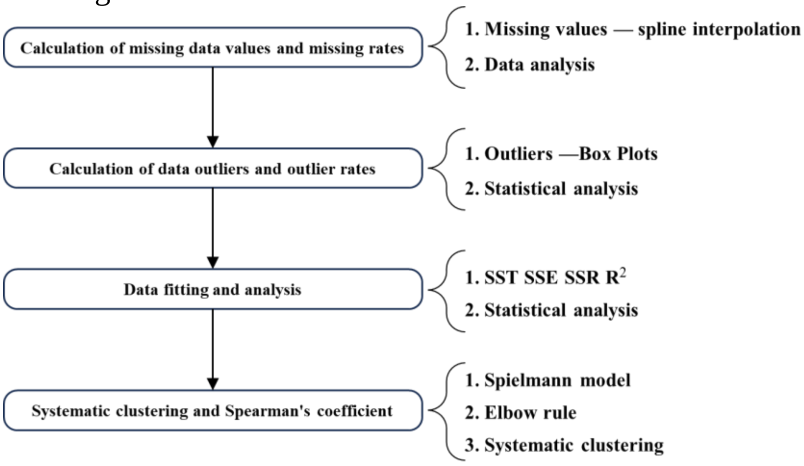

Calculation of missing data values and missing rates
1. Missing values — spline interpolation
2. Data analysis
Calculation of data outliers and outlier rates
1. Outliers —Box Plots
2. Statistical analysis
Data fitting and analysis
1. SST SSE SSR R2
2. Statistical analysis
Systematic clustering and Spearman's coefficient
1. Spielmann model
2. Elbow rule
3. Systematic clustering

Figure 1: Logic diagram.

First, the missing values are screened out and supplemented by means of interpolation algorithm for missing values, and box plots are used for outliers. Secondly, the products with very few sales (<0.1%) are screened out. Box plots are used to screen out the abnormal data and eliminate them, so as to ensure that the data are reliable and stable and facilitate the study of the logical relationship between the data. Finally, the data were analyzed and processed to obtain the distribution pattern of each vegetable category and single product, and the Spearman's correlation coefficient of each category was obtained by analyzing the sales by period and the Spearman's correlation coefficient of each product was obtained by using systematic clustering method to classify the single product into different categories and to find out the Spearman's correlation coefficient of the single product, as shown in Figure 1.

3. Discussion of results

3.1. Missing Data Values and Missing Rates

In the vegetable sales data, there are vegetable items with low three-year sales volume (<0.1%). There is a part of the value of the missing or abnormal phenomenon, the existence of the value will affect the accuracy of the final results. Screening out the missing values and abnormal values will ensure the accuracy of the analysis. The data for different vegetable species were processed to obtain the missing data as shown in Table 1.

Table 1: Missing values.

kind	cauliflower	philodendron	capsicum	eggplant	Edible mushroom	Aquatic rhizomes
Quantity	1750	2006	1860	1995	580	799
Missing rate	1.19	1.37	1.27	1.36	0.39	0.54

The data were organized to obtain the missing values and missing rates of the data and refined using spline interpolation. Processing six types of data found special cases: (a) some data in the statistics deviate from the neighbouring proximity of a large degree, the emergence of a sudden drop and rise, for the first outliers; (b) some data in the logic of the lack of correlation is too contradictory, for the second kind of outliers.

3.2. Data outliers and outlier rates

The t-test was used to test the data outliers one by one and screen out the abnormal data to improve the accuracy of the results. Box plot in Fig. 2 was used to process the data and screen the outliers.

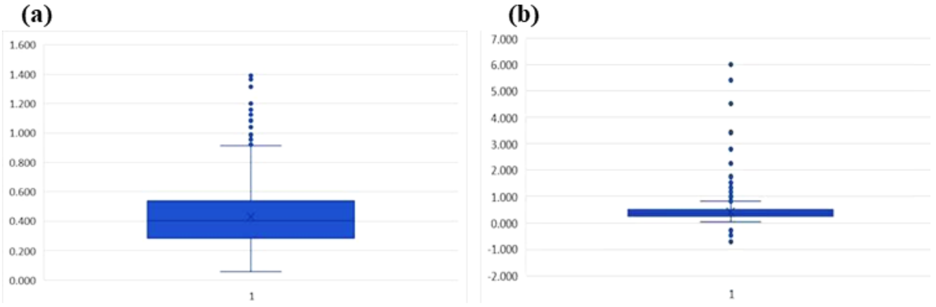


Figure 2: Box diagram of milky cabbage (a), Yunnan cabbage (b).

Vegetable data were processed using strict probability definitions to obtain the corresponding outlier data and calculate the corresponding outlier rate and remove the unqualified data in Table 2.

Table 2: Outliers.

kind	cauliflower	philodendron	capsicum	eggplant	Edible mushroom	Aquatic rhizomes
Quantity	994	1025	1120	1028	485	332
Anomaly rate	0.68	0.70	0.76	0.70	0.33	0.22

3.3. Fitting and Data Analysis

We need to conducting data processing and linear fitting to study the commonality of the data. The changes in vegetable sales volume are analysed from the perspectives of daily, monthly, and yearly dimensions, obtaining a table. The regression fitting determines their distribution and trends.

After fitting the regression treatment, the functional relationship is obtained. Let the corresponding sample points be set to $(x_i, y_i), i = 1,2,...,n$. Then the fitted curves are set to 1 and 2 $y \ = kx \ + b$. The fitted values are set to be $y_i = kx_i + b$, Goodness-of-fit of R2,Overall sum of squares is SST. The sum of squared errors is SSE. The regression sum of squares is SSR. The goodness-of-fit satisfies the following conditions.

$$0 \leq R^2 = \frac{SSR}{SST} = \frac{SSR - SSE}{SST} = 1 - \frac{SSE}{SST} \leq 1 \tag{1}$$

The closer R2 is to 1 the closer the sum of squared errors is to 0, indicating a better fit as shown in Figure 3.

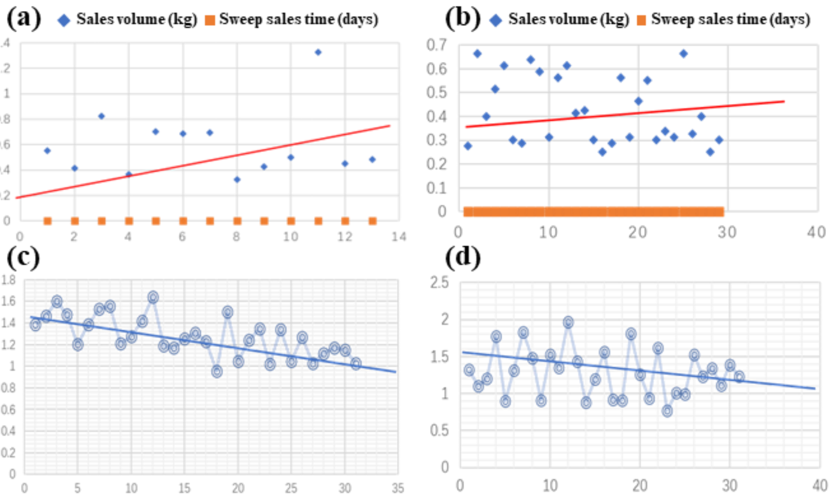


Figure 3: Daily sales of eggplant (a), choy sum (b) and monthly sales of Chinese cabbage (c), cabbage moss (d) in Dalung.

Analysis can be concluded: (1) Study on daily change of extraction by observing the daily sales volume of eggplant and Chinese cabbage. The data is fitted to the corresponding lines (a) (b). The overall performance was observed in a zigzag upward trend. Overall sales increased or remained basically stable. (2) Extracting the monthly change study, the corresponding straight line (c) (d) was obtained by observing the monthly sales volume of Chinese cabbage and cabbage moss and fitting the data to it. It was observed that the whole showed a zigzagging downward trend and the sales volume decreased as a whole.

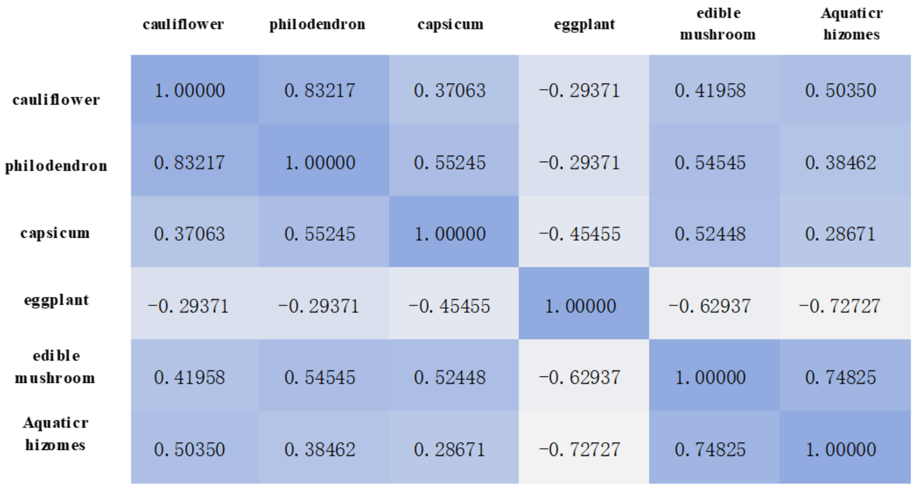
3.4. Systematic clustering and Spearman's coefficient

Spearman model is used to analyze the distribution law and mutual relationship between individual categories and product sales. At the same time, define two groups of basic data, X and Y. The relevant Spearman grade correlation coefficients are shown below 2:

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

where the numerator core number is the rank difference between the corresponding values of X and Y. The problem sample is large. For the statistic $r_s\sqrt{n-1} \sim N(0,1)$, $H_0 : r_s = 0, H_1 : r_s \neq 0$. By verifying $r_s\sqrt{n-1}$, a p-value of 0.05 relative to 0.05 can be derived.

After merging and preprocessing the data, the total sales for the quarter are calculated and the resulting matrix of total sales is brought into MATLAB and solved to obtain the Spearman's correlation coefficient matrix. Visualization of the correlation coefficient matrix results as shown in Figure 4 in the following Spearman's correlation coefficient plots for different categories.



	cauliflower	philodendron	capsicum	eggplant	edible mushroom	Aquaticr hizomes
cauliflower	1. 00000	0. 83217	0. 37063	−0. 29371	0. 41958	0. 50350
philodendron	0. 83217	1. 00000	0. 55245	−0. 29371	0. 54545	0. 38462
capsicum	0. 37063	0. 55245	1. 00000	−0. 45455	0. 52448	0. 28671
eggplant	−0. 29371	−0. 29371	−0. 45455	1. 00000	−0. 62937	−0. 72727
edible mushroom	0. 41958	0. 54545	0. 52448	−0. 62937	1. 00000	0. 74825
Aquaticr hizomes	0. 50350	0. 38462	0. 28671	−0. 72727	0. 74825	1. 00000

Figure 4: Spearman coefficient correlation analysis by category.

The analysis obtained conclusions: (1) The basic dishes are completely associated with themselves with a coefficient of 1. (2) The cauliflower and cauliflower, aquatic roots and edible mushrooms have the closest relationship with coefficients of 0.832 and 0.748. (3) cauliflower, cauliflower, as well as chili peppers are more highly correlated, and the edible mushrooms and aquatic roots and tubers are more highly correlated. This gives a picture of the correlation between individual items, which facilitates a more comprehensive understanding of the types of vegetables sold.

After the outliers were screened, there were still many items which made it difficult to observe the Spearman correlation coefficient matrix. Therefore, this paper systematically clusters the data given by the topic, reduces its complexity, and synthetically processes the data. Multiple clustering centres are initialized successively, and data objects are allocated to the nearest set. The final result is output after convergence or iteration. Eliminating the effect of different scales, the data in each column were normalized as follows.

$$V_{std} = \frac{V_m}{V_{\max}} \tag{3}$$

The normalized data were imported into SPSS for systematic clustering. The distortion degree of each category was used as the square sum of the distance between the center of the category and its internal members. Assuming that there were n samples in total, they were classified into K major categories, and the location of the center of the category was denoted as u_k.The total degree of distortion for all categories is obtained by referring to the following equation 4 to obtain the clustering coefficient *J*.

$$J = \sum_{k=1}^{K} \sum_{i \in C_k} \left| x_i - u_k{}^2 \right| \tag{4}$$

Also, the following table in Figure 5 is obtained.

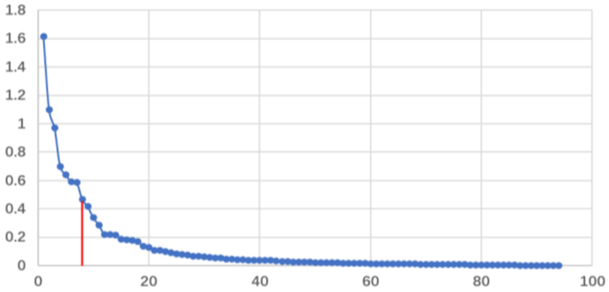
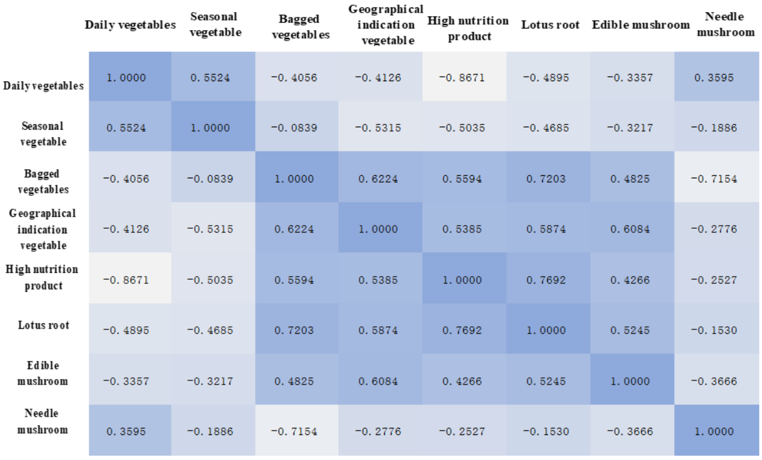


Figure 5: Elbow law coefficient diagram.

From the coefficient graph, we can get the conclusion: (1) According to the aggregation coefficient line graph can be analyzed to get: when the category is 8, the downward trend of the line is slow and when the category value is set to 8, the change of the degree of aberration is the largest. (2) The aberration value K value from 1 to 8, the aberration degree changes dramatically, more than 8 after the curve becomes flatter. Therefore, the elbow is K = 8. We can analyze the data in Figure 6 to get by using Spearman's criterion to.



	Daily vegetables	Seasonal vegetable	Bagged vegetables	Geographical indication vegetable	High nutrition product	Lotus root	Edible mushroom	Needle mushroom
Daily vegetables	1.0000	0.5524	−0.4056	−0.4126	−0.8671	−0.4895	−0.3357	0.3595
Seasonal vegetable	0.5524	1.0000	−0.0839	−0.5315	−0.5035	−0.4685	−0.3217	−0.1886
Bagged vegetables	−0.4056	−0.0839	1.0000	0.6224	0.5594	0.7203	0.4825	−0.7154
Geographical indication vegetable	−0.4126	−0.5315	0.6224	1.0000	0.5385	0.5874	0.6084	−0.2776
High nutrition product	−0.8671	−0.5035	0.5594	0.5385	1.0000	0.7692	0.4266	−0.2527
Lotus root	−0.4895	−0.4685	0.7203	0.5874	0.7692	1.0000	0.5245	−0.1530
Edible mushroom	−0.3357	−0.3217	0.4825	0.6084	0.4266	0.5245	1.0000	−0.3666
Needle mushroom	0.3595	−0.1886	−0.7154	−0.2776	−0.2527	−0.1530	−0.3666	1.0000

Figure 6: Spearman correlation coefficient analysis of simplified categories.

Analysis to get the conclusion: (1) The basic dishes and its own coefficient of 1, completely associated. (2) Daily green vegetables and seasonal products have a greater degree of correlation, which also has a greater relationship with the seasonal turnover of vegetables in the market. It can be analyzed to the seasonal changes in vegetables and the sale of vegetables.

In summary, a set of simplified strategies for data preprocessing, statistical analysis and correlation analysis of vegetable supply and marketing type problems were developed. First, preprocessing was required to screen out and supplement sales due to the presence of anomalies and missing data. Second, the relationship between the individual vegetable items in each category over time is investigated. Finally, the data vegetable categories are simplified using clustering. The linear relationship of vegetable sales over time is obtained to be strong, and 8 types of vegetables can be simplified to 3 types using clustering. The future sales of vegetables are analyzed from three aspects of region, nutritional value and season, so as to simplify the problems of vegetable supply and marketing.

4. Dissemination and application of models

There are two advantages of the model. Firstly, the model is embedded layer by layer and checked hierarchically, and each step utilizes a large number of test formulas to repeatedly verify the accuracy and maximize the correctness and reasonableness of the results. Secondly, the model simplifies the handling of the problem. For situations involving many variables, the system simplifies the factors by clustering and handles the problem systematically, making it easy to apply and handle the problem in the field. There are two shortcomings of the model: To begin with, the model involves the market problem, the model to improve the yield is more complex, resulting in its analysis process is cumbersome and time-consuming; for another, vegetable sales category is more complex, the reality of the existence of many factors affecting.

The variety of vegetable sales and the complexity of influencing factors make the study of vegetable sales extremely challenging. This article starts from reality and investigates the inherent connections through screening, fitting, and analysis, laying the foundation for future models regarding vegetable sales volume, pricing, and other issues. In the future, it can be well applied to solving market sales issues, travel promotion issues, factory development issues, etc., with certain promotional value.

References

[1] Santamaria P, Signore A. How has the consistency of the common catalogue of varieties of vegetable species changed in the last ten years [J]. Scientia Horticulturae, 2021, 277: 109805.

[2] Rahman S M E, Mele M A, Lee Y T, et al. Consumer preference, quality, and safety of organic and conventional fresh fruits, vegetables, and cereals[J]. Foods, 2021, 10(1): 105.

[3] Asgari S D, Mohammadi E, Makui A, et al. Data-Driven Robust Optimization Based on Position-Regulated Support Vector Clustering[J]. Journal of Computational Science, 2024: 102210.

[4] Shen Q, Qiu Y. A Novel Text Ensemble Clustering Based on Weighted Entropy Filtering Model[C]//Journal of Physics: Conference Series. IOP Publishing, 2021, 2024(1): 012045.

[5] Rodriguez-Pretelin A, Nowak W, Morales-Casique E. Optimization-based clustering of random fields for computationally efficient and goal-oriented uncertainty quantification: Concept and demonstration for delineation of wellhead protection areas in transient aquifers[J]. Advances in Water Resources, 2022, 162: 104146.

[6] Zhou J, Lai Z, Miao D, et al. Multigranulation rough-fuzzy clustering based on shadowed sets [J]. Information Sciences, 2020, 507: 553-573.