ISSN 2616-5775 Vol. 6, Issue 13: 99-106, DOI: 10.25236/AJCIS.2023.061315

Research on Vegetable Pricing Strategy of Supermarket Based on Statistical Analysis and Optimization Algorithm

Xinpeng Yu[1], Xiaoqiao Qin[2], Yasong Cao[3], Ninghui Wu[3], Yaqi Tu[2], Ziyang Yu[2,*]

[1]*School of Electronic Engineering, Guilin Institute of Information Technology, Guilin, China*
[2]*School of Information Engineering, Guilin Institute of Information Technology, Guilin, China*
[3]*School of Mechanical and Electrical Engineering, Guilin Institute of Information Technology, Guilin, China*
[*]*Corresponding author: 2076781185@qq.com*

Abstract: *With the development of social economy, consumers' demand for the quality of vegetables is increasing, and the quality of vegetable goods changes over time. In this paper, we address the issue of the stocking volume and pricing strategy of vegetables in superstores and adopt statistical methods and programming language for data preprocessing, including data searching, cleaning, transforming, integrating and statute, aiming to optimize the pricing of vegetables. First, this study analyzes the relationship between vegetable category and single product sales and time through Pearson coefficient to reveal the sales distribution patterns and interrelationships. Second, a mathematical model is used to predict the restocking volume and pricing strategy of the superstore in the coming week, and the revenue is maximized through model diagnosis, fitting function selection and optimization of the SLSQP algorithm. Finally, the Northern Eagle optimization algorithm is used to screen and optimize the price of saleable individual items in combination with the restocking volume and pricing constraints to improve the profit of the superstore.*

Keywords: *Vegetable pricing strategy; Data preprocessing; Pearson's coefficient; Mathematical model; Optimization algorithm*

1. Introduction

In contemporary fresh produce superstore operations, the freshness of vegetables, as an important consumer attraction, challenges the superstore's replenishment decisions and pricing strategies. Supermarkets need to constantly adjust the vegetable replenishment volume in the absence of accurate information on individual products and data on the cost of incoming goods. By constructing a mathematical model, this paper aims to analyze the correlation between vegetable sales volume and time in order to reduce replenishment blindness and improve profits. This paper poses three problems: first, building a model to reveal the distribution pattern of vegetable sales and the relationship between different varieties; second, analyzing the relationship between total vegetable sales and profit, and predicting the future replenishment volume to formulate an effective pricing strategy; and finally, considering controlling the number of individual items and the ordering volume within confidence intervals to optimize the replenishment volume and the pricing strategy, to achieve the supply-side structural reform and to improve the profit of the superstores.

2. Relevance analysis

2.1 Total Vegetable Category Sales vs. Time

In the data analysis in this paper, individual product names were mapped to their numbers, which were prioritized and retained for subsequent analysis. Through data integration methodology, statistics were summarized for the vegetable category, covering a large amount of vegetable sales data. The time span analyzed in this paper extends from July 1, 2020, to June 30, 2023, a total of 1095 days over three years. Utilizing specific software, this paper provides an overall summary of individual vegetable products and vegetable categories versus specific time periods.

ISSN 2616-5775 Vol. 6, Issue 13: 99-106, DOI: 10.25236/AJCIS.2023.061315

Regarding the total vegetable category sales versus time, this paper shows the different vegetable categories and their sales through the summarized data. For example, the flower and leafy vegetables, numbered 1011010101, had a total sales volume of 213524.167 units; cauliflower vegetables, numbered 1011010201, had a total sales volume of 3,922.293 units; aquatic roots and tubers, numbered 1011010402, had a sales volume of 23,985.026 units; eggplant vegetables, numbered 1011010501, had a sales volume were 35,239.753 units; peppers, No. 1011010504, with sales of 70,248.43 units; and edible mushrooms, No. 1011010801, with total sales of 124,056.249 units. These data provide important basic information for subsequent analysis.

In the analysis of this paper, the Pearson correlation coefficient method was applied to explore the linear correlation between time and sales volume of each category of vegetables. Pearson's correlation coefficient is a key indicator of the degree of linear correlation between two variables, and its value ranges from -1 to 1. The closer the correlation coefficient is to 1, the stronger the linear correlation between the two variables; if the correlation coefficient is close to -1, the weaker their linear correlation. The judgment of correlation is based on the value of correlation coefficient, which can be categorized from very strong correlation to very weak correlation or no linear correlation [1].

In order to analyze the relationship between each category of vegetables and the sales volume of a single product, this paper first analyzes the relationship between time and vegetable categories. The relationship between total sales and time is explored by using the Pearson's coefficient method, which takes time as the independent variable and vegetable category sales as the dependent variable and calculates the Pearson's correlation coefficient between them. The calculation of Pearson's correlation coefficient involves the difference of each sample data from its mean value and the ratio of the sum of the products of these differences to the open square of their respective sums of squares.

On this basis, each vegetable subcategory was separately grouped into vegetable major categories. The specific model diagram is shown in Figure 1:

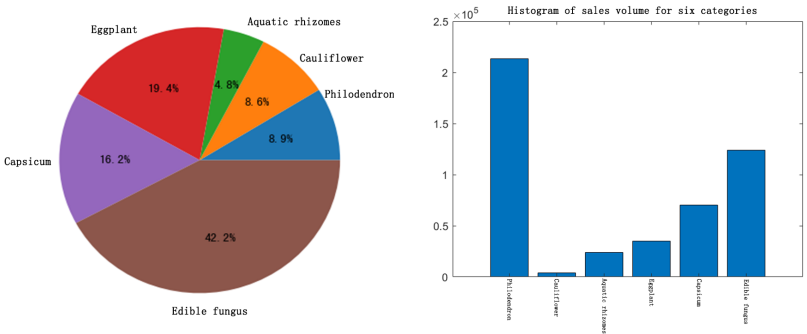

Figure 1: Statistical Modeling of Vegetable Broad Categories

The analysis reveals that flowering and leafy vegetables accounted for the largest share of the market, which indicates the prominence of their sales in the market. In contrast, the market share of cauliflower vegetables is relatively small, while the market share of aquatic root vegetables has increased. The market share of eggplant vegetables is slightly higher than aquatic root vegetables but lower than edible mushrooms.

Numerous factors influence the sales of vegetable categories, including personal preferences, vegetable quality, price, supply chain issues and nutritional value. Through a step-by-step analysis of a wide range of vegetable categories, this paper explores the reasons for the near-exclusive market share of leafy and flowering vegetables. Foliar vegetables include numerous individual varieties, such as spinach, oleander, and lettuce, which are tasty, nutritious, and affordable. In addition, leafy and flowering vegetables have a wide range of maturity periods and are adaptable to a variety of seasons, making it easy for consumers to access a diverse range of choices.

In contrast, cauliflower vegetables have a smaller share of the market. This phenomenon is viewed from a practical point of view, as cauliflower vegetables have fewer varieties and tend to be concentrated in specific seasons compared to leafy vegetables, resulting in their lower popularity in the market. Combining these factors, it is reasonable that cauliflower vegetables dominate the market.

On this basis, based on the Pearson correlation coefficient method using MATLAB software to establish a mathematical model, the model will be three years of data volume is divided into 1095 days, the daily total sales of vegetable categories into the model, the statistics of three years of the trend of

ISSN 2616-5775 Vol. 6, Issue 13: 99-106, DOI: 10.25236/AJCIS.2023.061315

vegetable sales and reflected in the model. The curve of the total sales volume of each category over time is shown in Figure 2 below.

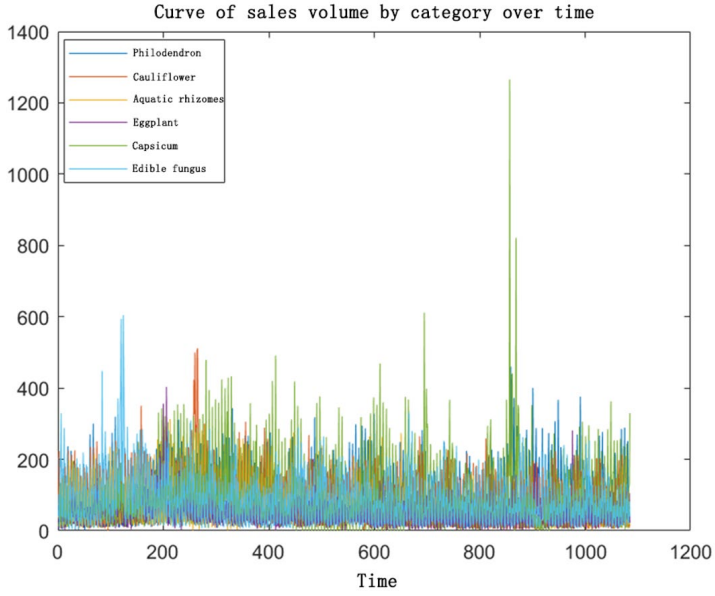

Figure 2: Curves for each category over time

Sales of pepper vegetables were volatile, topping out sharply due to promotional activity and imbalances between supply and demand. Edible mushroom sales were stable but showed seasonal increases, while eggplant sales fluctuated with spring maturity. Leafy and flowering vegetables, while generally stable, saw a spike in sales that may be related to seasonal trends such as winter cabbage purchases. Aquatic root vegetables maintain steady sales due to their nutritional value, and cauliflower sales spikes correspond to seasonal events such as spring canola bloom.

2.2 Heat map to solve the distribution pattern and interrelationship of the sales volume of each category of vegetables.

Heat map is also known as correlation coefficient diagram, this paper analyzes the distribution law of each vegetable category and sales volume: according to the size of the correlation coefficient corresponding to the different color squares in the heat map, we can judge the degree of correlation between the variables. In this paper, the independent variable is time X, and the dependent variable is the total sales volume of single product Y2, and the correlation coefficient between the variables is calculated by the formula:

$$\rho_{x1x2} = \frac{Cov(X,Y2)}{\sqrt{DX_1,DX_2}} = \frac{EX_1X_2 - EX_1 * EX}{\sqrt{DX_1 * DX_2}} \quad (1)$$

Where ρ represents the correlation coefficient, Cov represents the covariance in, and E represents the mathematical expectation or mean. From the above Pearson correlation coefficient method, the higher the correlation coefficient, the higher the degree of linear correlation between the variables. The degree of linear correlation of the heat map can be determined by the size of the correlation coefficient ρ value. In this paper, based on the characteristics of the heat map with the determination of the degree of linear correlation, the distribution pattern and interrelationship of the sales volume of each category of vegetables are presented in the heat map to realize the distribution pattern and interrelationship of the sales volume of each category of vegetables, and the curve of the above-mentioned categories with the change of time is further visualized. Thus, the regular relationship that exists between vegetable categories and total sales volume is presented in a concrete form. The heat map presenting the regular relationship between vegetable categories and total sales volume is shown in Figure 3 below [2].

ISSN 2616-5775 Vol. 6, Issue 13: 99-106, DOI: 10.25236/AJCIS.2023.061315

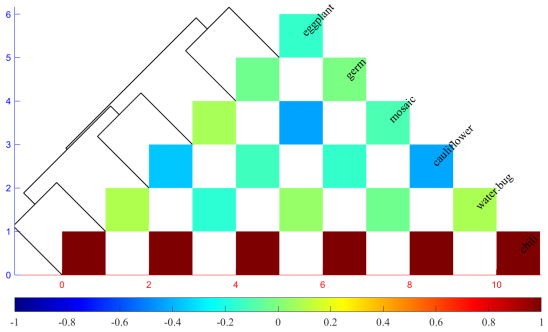

Figure 3: Heat map of the regular relationship between vegetable category and total sales volume

3. Category replenishment and pricing analysis

3.1 Model Preparation and Preprocessing

In this paper, an autoregressive moving average model (ARIMA) is used to predict the daily replenishment demand for various vegetable categories in the coming week. The model is defined by three parameters: the autoregressive coefficient, the order of differencing, and the moving average coefficient. These parameters represent the relationship of the variable to its past values, the order of difference required for data smoothing, and the lagged effect under error, respectively.

Prior to ARIMA model construction, thorough data preprocessing is essential to ensure the accuracy and validity of the model. This includes checking and filling in missing values with the mean of the dataset, defining the raw data, and creating independent time series. Trends in the data are assessed using scatter plots and non-smooth series are processed until autocorrelation and partial correlation functions show non-significant or non-zero values.

Appropriate time series models are then developed based on the identified patterns, and key parameter estimation is performed to make accurate model predictions. Finally, the established hypotheses are tested for statistical significance and predictive power.

This meticulous approach aims to develop an accurate and reliable ARIMA model to help supermarkets in effective inventory management and profit optimization [3, 4].

3.2 Solving the model

To investigate the relationship between total sales and cost-plus pricing for vegetable categories, a statistical model was fitted and analyzed. Each of the six different vegetable categories was subjected to model diagnostics, including histogram plus estimated density plot tests and other tests of correlation plots to assess the stability of the model.

After these tests, the models used showed strong stability. Subsequently, the paper applies linear, logarithmic, and power functions to fit the data. In this analysis, cost-plus pricing is treated as the unit price of goods sold. Model plots of the relationship between total sales volume and cost-plus pricing are shown for six different vegetable categories.

Through these models, this paper successfully explores the interrelationship between sales volume and cost-plus pricing for vegetable categories. This analysis contributes to a deeper understanding of how pricing strategy affects sales volume, which in turn provides data support for superstores to optimize profits. The methodology does not only consider the impact of price on sales volume, but also focuses on the potential differences between categories, thus providing an important basis for developing more accurate pricing strategies.

ISSN 2616-5775 Vol. 6, Issue 13: 99-106, DOI: 10.25236/AJCIS.2023.061315

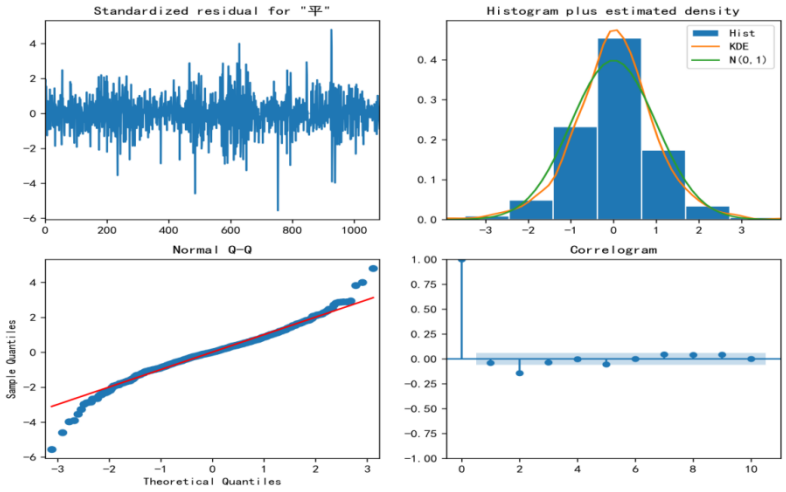

Figure 4: Diagnostics of ARIMA model for foliar species

As shown in Figure 4, this figure shows the model diagnosis of foliage vegetables, and the standardized residual test of foliage vegetables can be seen from the model, and the residual can be used as the observation value of the error. It can be seen that because of the foliage vegetables sell well, so the model has a large number of points and conforms to the above law; from the histogram plus the estimated density plot test model can be seen, the curve is steeper than the (0,1) normal distribution curve, the center point has exceeded the highest point of the model; its test verifies that the curve in addition to a small portion of the points of deviation, basically a straight line; due to the exception of some of the small deviation of the value of the points is almost a straight line Since the points are almost in a straight line except for some very small deviations, the data in this group are also in a straight line and converge under the correlation test schematic diagram. In summary, the results of the model diagnosis of leafy and flowering vegetables show that the stability of the model made from the sales data of leafy and flowering vegetables is good[5].

Using Python programming language, by constructing the relationship model between sales volume and cost-plus pricing (unit price of sales) were fitted with linear, logarithmic, and power functions respectively to obtain the corresponding function fitting goodness of fit for each vegetable category, as shown in Table 1:

Table 1: Goodness of fit of functions for each vegetable category

Category	Linear function	Logarithmic function	Power function
Aquatic Roots	0.11745080069648461	0.12923451930385366	0
Foliage	0.044656284163110516	0.06654722980518557	0
Cauliflower	0.10025714280460707	0	0
Eggplant	0.03763208851021638	0.05244971750068983	0
Peppers	0.02222875889460274	0.06447516978371415	-2.2204463e-16
Mushrooms	0.025661297021104756	0	0

According to the corresponding function fitting superiority of each vegetable category obtained, the function with higher fitting superiority was preferentially selected to establish the model, and the function fitting results of each vegetable category were obtained, as shown in Table 2 below:

Table 2: Function fitting results for each vegetable category

Sequences	Classification	Model Name	Model parameter
1	Aquatic Roots	Logarithmic	-26.1101487, 2.517253347.83. 6..
2	Flower and Leaf	Logarithmic	-36.00775817714, 2.354351525, 209....
3	Cauliflower	Linear	-2.8916556593, 64.2045166
4	Eggplant	Logarithmic	-8.16177241281, 1.62367181, 36.470...
5	Pepper	Logarithmic	-17.07560417, 3.10816626, 105. 16..
6	Mushrooms	Linear	-3.2887825504, 92.53757998

Using the SLSQP (Sequential Least Squares Programming) algorithm in mathematical optimization algorithms, the predicted prices for each vegetable category can be obtained by iteratively approximating the optimal solution of the objective function and constraints.

ISSN 2616-5775 Vol. 6, Issue 13: 99-106, DOI: 10.25236/AJCIS.2023.061315

4. Individual replenishment and pricing analysis

To solve the problem of optimizing the number of single items ordered and the total number of single items available for sale in the vegetable category, this paper employs the Northern Eagle optimization algorithm, which aims to maximize the profitability of the superstore while meeting specific requirements. This approach first focuses on the quantity of saleable items to be stocked in a specified period, by focusing on the listed individual items to be screened and retained to determine the appearing individual items and their correspondences.

In the application of the algorithm, the minimum value of the number of single items to be displayed is set as a fixed value. Incoming quantities below this value are extracted, and these data are processed to lay the foundation for subsequent replenishment volume prediction and price analysis. After completing the measurement of the replenishment volume, the data is further processed to filter out the data below the set value and retain the original value for the data greater than this value. In terms of pricing strategy, if the selling price is lower than the inlet price, the inlet price is used as the selling price, thus promoting the maximization of superstore profits.

With this approach, this paper effectively balances the needs of superstores between maintaining inventory and optimizing revenue, while also considering the impact of market dynamics and consumer behavior. The application of the Northern Pale Eagle optimization algorithm shows its unique advantages in dealing with complex supply chain and inventory management problems and provides an innovative solution for maximizing superstore profits.

The Northern Pale Eagle algorithm has the following phases:

(1) Initialization phase

The members of the population need to be initialized randomly in the search space before the algorithm starts, the formula can be as follows.

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_i \\ \vdots \\ X_N \end{bmatrix}_{N\times m} = \begin{bmatrix} x_{1,1} \cdots x_{i,j\cdots} \\ \vdots \\ x_{i,1} \cdots x_{i,j} \cdots \\ \vdots \\ x_{N,1} \cdots x_{N,j} \cdots \end{bmatrix} \tag{2}$$

where X denotes the population matrix of the northern goshawk, X_i is solved from the ith individual, N is the number of population members, and m is the dimensionality of the problem space.

The objective function values for the goshawk population can be used in an objective function value vector representation as shown in the following equation:

$$F(x) = \begin{bmatrix} F1 = F(X_1) \\ F_i = F(X_2) \\ F_N = F(X_N) \end{bmatrix}_{N\times 1} \tag{3}$$

where F is the vector for obtaining the objective function and F_i is the value of the objective function obtained for the ith solution.

(1) Phase I: Prey identification

The hawk randomly selects a prey in the first phase of the hunt and attacks quickly. By the uncertain selection of prey in the search space, this phase increases the NGO search capability. The purpose of global search is to identify the optimal system:

$$P_i =, X_k, i = 1,2,\cdots,N, k = 1,2,\cdots,i-1,i \tag{4}$$

$$X_{i,j}^{new,P1} = \begin{cases} x_{i,j} + r(p_{i,j} - Ix_{i,j}), Fp_i < F_i \\ x_{i,j} + r(x_{i,j} - p_{i,j}), Fp_i \geq F_i \end{cases} \tag{5}$$

$$X_i = \begin{cases} X_i^{new,P1}, F_i^{new,Pi} < F_i \\ X_i, F_i^{new,P1} \geq Fi \end{cases} \tag{6}$$

where P_i is the position of the ith northern goshawk prey, F_{pi} is the value of its target as a function of its objective, k is a random integer in the interval $[1, N]$ that is not i, and I has a value of 1 or 2. Parameters r and l are random numbers used for searching as well as for dynamically updating the

ISSN 2616-5775 Vol. 6, Issue 13: 99-106, DOI: 10.25236/AJCIS.2023.061315

behavior of the randomly generated NGO.

(2) Phase 3: Escape state

When attacking a prey, the prey tries to escape. In the process of the hawk trailing the prey at high speed, this simulation algorithm is to optimize and upgrade the ability to utilize the algorithm's search space locally. In the NGO algorithm assumes the attack position with radius R. Mathematical modeling is performed and the formula is shown below:

$$x_{i,j}^{new,P2} = x_{i,j} + R(2r-1)x_{i,j} \tag{7}$$

$$R = 0.02(1-\frac{t}{T}) \tag{8}$$

$$X_i = \begin{cases} X_i^{new,P2}, F_i^{new,P2} < F_i \\ X_i, F_i^{newc,P2} \geq F_i \end{cases} \tag{9}$$

In this paper, we use the Northern Eagle algorithm to build a mathematical model with Matlab as shown in Figure 5:

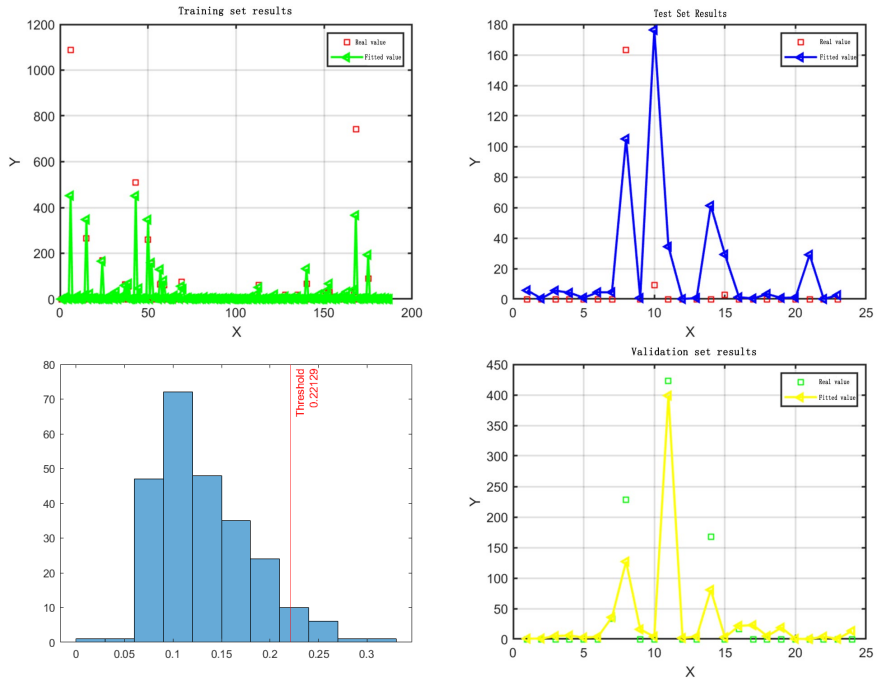

Figure 5: Results between forecast sales

5. Conclusions

This study excels in data processing, notably in comprehensive data preprocessing and outlier detection, which ensures the foundational accuracy and reliability of the analysis. It maintains data integrity by minimizing alterations, allowing for deeper insights. The use of the Pearson correlation coefficient method simplifies correlation analysis, adding visual clarity to the relationship between variables. However, the study faces challenges in parameter selection and handling complex or unexpected data trends, suggesting a need for future model optimization and methodological advancements. Overall, it demonstrates high accuracy and reliability but requires improvement in managing more intricate data scenarios.

References

[1] Zhao Yao; Yu Lijuan; Su Yixin; Zheng Tuo; Tong Guangbo. Cleaning and de-duplication of grid load data based on cluster analysis and Pearson correlation coefficient method [J]. Ship Power Technology, 2023, 43(06):69-75. DOI:10.13632/j.meee.2023.06.019

[2] Guo, L.; Guo, Z. X.; Jia, H. T.; Fan, R. Y. Identification of residential electricity theft based on

ISSN 2616-5775 Vol. 6, Issue 13: 99-106, DOI: 10.25236/AJCIS.2023.061315

Pearson correlation coefficient and SVM [J]. Journal of Hebei University (Natural Science Edition), 2023, 43(04):357-363.

[3] GUO Zuoxiang; LIU Dongpeng. Evaluation of hand-foot-mouth disease incidence in Zhangye City based on autoregressive moving average model [J]. Chinese Journal of Viral Diseases, 2023, 13(05):390-394. DOI:10.16505/j.2095-0136.2023.5013

[4] Jing Wang; Miaomiao He; Jian Ding; Yonghua Li. A spatio-temporal graph convolutional network for multidimensional time series anomaly detection [J/OL]. Journal of Xi'an University of Electronic Science and Technology, 1-11[2023-11-21] https://doi.org/10.19665/j.issn1001-2400.20230804.

[5] An Yuyuan; Jia Chaowen; Hu Liuchun; Zhang Xueshuai; Liu Xiang; Yan Bo; Li Yanping. A high-resolution and high-precision direction finding method for virtual interferometer based on linear prediction model [J]. Electronic Information Countermeasures Technology, 2021, 36(02):14-17.