Economic Data Analysis Based on Data Mining

Hongpo Wang*, Shuai Shao, Lingyun Gao
e-mail: nudtwhp@126.com

Complex Aeronautical System Simulation Laboratory Beijing China

Abstract—With the development of equipment technology, the cost of equipment development is rising too. How to reasonably plan, manage and use the cost of equipment is the concern of equipment management departments and researchers. Data mining technology can find the necessary information from the equipment economic data, which is of great significance to improve the cost-effectiveness. In view of the difficulty in estimating the funds in the project demonstration stage, this paper first uses the correlation analysis method to study the correlation between the normalized eight cost data and the total cost, removes the cost items with weak correlation, and only retains the design person year ratio, material person year ratio, test person year ratio and the total cost person year ratio as the later research object. Next, through cluster analysis, it is found that the aggregation degree of material man year ratio and total cost man year ratio is the best, and the relationship is approximately linear. Then, the 1~5 order regression model is established by using the least square method. Finally, the regression model is verified based on the project data.

Keywords-Economic data; data mining; regression

1 INTRODUCTION

Economic data mining is to mine the hidden knowledge from the massive cost related data. With the rapid development of new technology, the complexity and informatization of equipment are increasing day by day, resulting in the rising cost of research and development. How to reasonably demonstrate and put forward the development cost in the initial stage of project development under the condition of extreme lack of information is the key concern of the arguer and decision-making body.

The cost of equipment scientific research project consists of pricing cost, revenue and contingency cost. The pricing cost includes design cost, material cost, outsourcing cost, special cost, test cost, fixed asset use cost (hereinafter referred to as fixed asset cost), salary cost and management cost [1] (the cost item of the new regulations has been adjusted. Since the data is collected according to the old regulations, the old regulations are used here). By collecting the economic data of equipment development and mining the relationship between various costs and total funds, we can correctly grasp the characteristics and laws of the composition of scientific research funds, which is of great significance to strengthen the management of equipment funds and improve the use efficiency of funds.

Economic data mining has been concerned by researchers. In order to maximize the benefit of equipment funds, wubaojun [2] and others have built an optimization model of equipment funds proportion based on combat capability, which provides a useful reference for improving the

FFIT 2022, October 28-30, Shenzhen, People's Republic of China

DOI 10.4108/eai.28-10-2022.2328441

benefit of equipment funds. Liaomeihong et al. [3] proposed a ship equipment maintenance cost estimation model based on data mining to improve the accuracy of ship equipment maintenance cost and reduce the estimation deviation. Zhangshouyu et al. [4] analyzed the budget demand of equipment maintenance funds based on data mining technology, determined the unreasonable budget demand, and put forward the idea of using data mining technology to solve complex data processing problems. Pan Jianguo et al. [5] used the grey theory to analyze the composition of the use support cost of a certain type of aircraft, determine the main influencing factors, and provide reference for further improving the use efficiency of equipment funds. Wanghongpo et al. [6] used the correlation analysis method to conduct correlation analysis on the eight costs of the project, such as scientific research personnel year, design cost and material cost, and gave the correlation between the scientific research personnel year and the eight costs, as well as the eight costs.

Aiming at the problem of equipment development cost estimation, this paper adopts data mining methods such as cluster analysis and regression modeling to analyze the economic data. Firstly, eight cost data are normalized according to the annual investment of the developer, and the correlation among different cost items is analyzed. Then, based on the correlation between parameters, the data are normalized transformation and cluster analysis, and the parameter characteristics are observed. Then, according to the data characteristics, the regression model is established based on the least square method, and the rationality of the model is analyzed. Finally, the whole paper is summarized.

2 DATA CORRELATION

Based on the economic data in reference [6], the cost data was normalized according to the annual investment of scientific researchers (i.e., the cost and total cost of the eight items were divided by the annual investment of scientific researchers). Spearman correlation analysis method was used to calculate the correlation of the eight cost data, and the correlation coefficients were obtained as shown in Table 1.

Hypothesis test data tables in reference [6], the correlation between DPYR, TPYR, MPYR, MPYR, APIR and TCPY is stronger. While the correlation between FAPYR, WPYR and TCPY is weak. There isn't connection between OPYR, DDPYR and TCPY.

According to reference [1], although the correlation between MPYR, APIR and TCPY is strong. However, according to the proportion of cost, fixed assets, wages, outsourcing, special and total cost correlation is weak. This paper focuses on the analysis of the relationship between DPYR, MPYR, TPYR and TCPY.

Table 1 Table of relative numbers of each item

ITEM	DPYR	MPYR	OPYR	DDPYR	TPYR	FAPYR	WPYR	MYR	APIR	TCPY
DPYR	1	0.5730	0.0068	0.1549	0.3026	0.3190	0.2954	0.5136	0.6144	0.6037
MPYR		1	0.0839	0.1335	0.4511	0.3307	0.1368	0.6634	0.8751	0.7858
OPYR			1	0.0297	0.2742	0.1070	0.0047	0.1564	0.0202	0.2197
DDPYR				1	0.1766	0.0600	0.3005	0.0985	0.0324	0.1207

TPYR					1	0.0506	0.1576	0.6295	0.5974	0.6688
FAPYR						1	0.1942	0.0793	0.2882	0.2695
WPYR							1	0.3000	0.2413	0.3277
MYR								1	0.7321	0.8813
APIR									1	0.8452
TCPY										1

Note: The full name of abbreviations.

Abbreviations	The full name	Abbreviations	The full name	Abbreviations	The full name
DPYR	Design person year ratio	TPYR	Test person year ratio	APIR	Annual person income ratio
MPYR	Material person year ratio	FAPYR	Person year ratio of fixed assets	TCPY	Total cost per person year
OPYR	Outsourcing person year ratio	WPYR	Wage person year ratio		
DDPYR	Dedicated person year ratio	MPYR	Manage person year ratio		

3 CLUSTERING ANALYSIS

Using data mining method to analyze economic data, we should: 1) analyze data correlation, get the correlation between each item cost and total cost, 2) transform the data and cluster analysis, get the conclusion that there is roughly linear relationship between material cost and total cost; 3) The regression function of material cost and total cost was obtained by least square fitting using cluster center parameters; 4) Formula verification; 5) Conclusion.

Cluster analysis is a method of grouping samples according to data similarity without given classification. It is an unsupervised learning algorithm. The input of clustering is a group of unlabeled samples, which are divided into several groups according to the distance or similarity of the data themselves. The principle of partitioning is to minimize the samples within the group and maximize the distance between the groups.

The K-means clustering algorithm divides the data into a predetermined number of classes K on the basis of minimizing the error function, and adopts distance as the similarity indicator, that is, the closer the distance between two objects is, the greater the similarity is. The algorithm process is as follows.

(1) Randomly select K objects from N sample data as the initial clustering center;

(2) Calculate the distance between each sample and each cluster center, and assign the objects to the nearest cluster;

(3) After all objects are allocated, the K cluster centers are recalculated;

(4) Compared with the K cluster centers obtained by the previous calculation, if the cluster centers change, go to step (2); otherwise, go to Step (4);

(5) Stop and output clustering results when the centroid does not change.

The result of clustering may depend on the random selection of the initial cluster center, which may cause the result to deviate seriously from the global optimal classification. In practice, in order to get better results, k-means algorithm is usually run multiple times with different initial clustering centers.

In order to study the data characteristics, DPYR and TCPY, MPYR and TCPY, TPYR and TCPY three clusters are analyzed. Since outsourcing and special charges are not related to other items of expenses, they are deducted from the total appropriation. In order to compare the distribution of data, the minimum-maximum normalization is adopted to transform the data.

Minimum-maximum normalization, also known as deviation normalization, is a linear transformation of raw data that maps numeric values to the interval [0,1]. The conversion formula is as follows:

$$x^* = \frac{x-min}{max-min} \tag{1}$$

Max is the maximum value of the sample data, min is the minimum value of the sample data, and max-min is the range. Deviation standardization preserves the relationships existing in the original data and is the simplest way to eliminate the influence of dimension and data range. The comparison of the results of clustering calculation is shown in Table 2 and Figure 1. ⊕ in the figure is the center of gravity of clustering, and different colors represent different clusters.

Table 2 Comparison table of cluster calculation

(1)

	DPYR	
	The mean distance between each data point and the center	***Percentage of clustering data volume***
Cluster 1	0.0602	67.61%
Cluster 2	0.0124	22.53%
Cluster 3	0.0603	9.86%
Average cluster distance was calculated	0.0494	

(2)

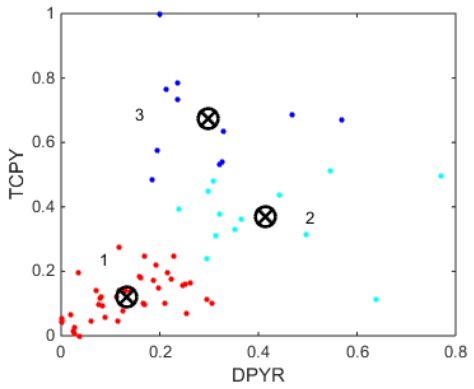
	MPYR	
	The mean distance between each data point and the center	***Percentage of clustering data volume***
Cluster 1	0.0060	64.79%
Cluster 2	0.0132	19.72%

Cluster 3	0.0402	15.49%
Average cluster distance was calculated	0.0127	

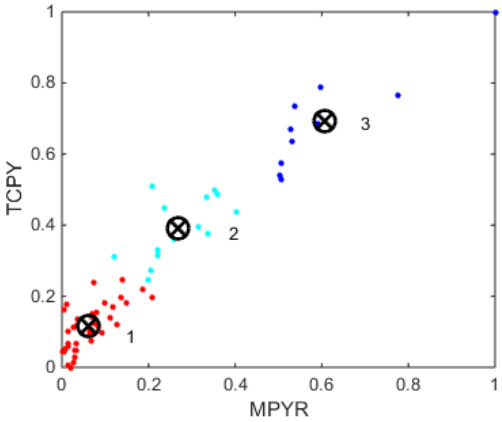
(3)

	TPYR	
	The mean distance between each data point and the center	***Percentage of clustering data volume***
Cluster 1	0.0572	67.61%
Cluster 2	0.0193	26.76%
Cluster 3	0.0401	5.63%
Average cluster distance was calculated	0.0461	

(1)



(2)



(3)

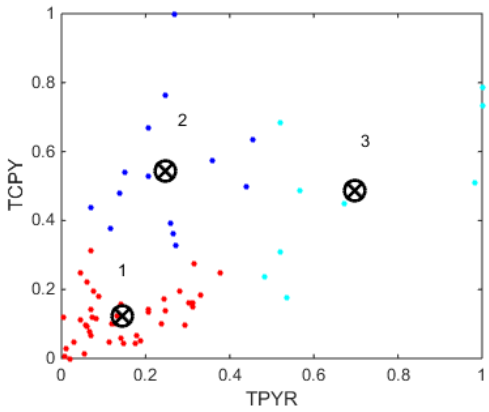


Figure 1. Clustering figure of DPYR and TCPY, MPYR and TCPY, TPYR and TCPY.

It can be seen from Table 2 and Figure 1 that the average distance between MPYR and TCPY is the lowest, and the aggregation degree of the data is the highest, with an approximate linear relationship, which can be used to build a regression model.

4 REGRESSION MODELING

The MPYR and TCPY are approximately linear, and the relationship between them can be studied by regression method.

Regression analysis is a method to analyze the causal relationship between dependent variables and independent variables. In general, there is no linear relationship between dependent variable Y, independent variable X and unknown parameters. The nonlinear relationship can be

transformed into an equivalent or approximate linear model through simple function transformation, and then solved by regression.

Suppose n sets of sample values are taken from the population (x_1, y_1), (x_2, y_2), …, (x_n, y_n). For n points in the plane, an infinite number of curves can be used to fit. The sample regression function is required to fit this set of values as well as possible.

The regression model chosen by ordinary least square method should minimize the sum of squares of residuals of all observations. The squared loss function is used.

The regression model is as follows:

$$\begin{aligned} y_i &= \overline{\beta_0} + \overline{\beta_1} x_i + \cdots + \overline{\beta_m} x_i^m + e_i \\ e_i &= y_i - \left(\overline{\beta_0} + \overline{\beta_1} x_i + \cdots + \overline{\beta_m} x_i^m \right) \end{aligned} \tag{2}$$

Where e_i is the error of sample (x_i, y_i) and m is the order of the equation.

Square loss function (i.e. sum of squares of residuals of the fitting curve) is

$$Q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \overline{y}_i \right)^2 = \sum_{i=1}^{n} \left(y_i - \overline{\beta_0} - \overline{\beta_1} x_i - \cdots - \overline{\beta_m} x_i^m \right)^2 \tag{3}$$

Fitting the curves by minimum determination of Q, that is $\overline{\beta_m}$, treating them as functions of Q, becomes a problem of taking an extreme, which can be obtained by taking a derivative. The partial derivative of the parameter to be estimated Q is obtained.

This is the solution of the least square method, obtaining the extreme point of the square loss function.

$$\begin{cases} \dfrac{\partial Q}{\partial \overline{\beta_0}} = \sum_{i=1}^{n} \left(y_i - \overline{\beta_0} - \overline{\beta_1} x_i - \cdots - \overline{\beta_m} x_i^m \right) = 0 \\ \dfrac{\partial Q}{\partial \overline{\beta_1}} = \sum_{i=1}^{n} \left(y_i - \overline{\beta_0} - \overline{\beta_1} x_i - \cdots - \overline{\beta_m} x_i^m \right) x_i = 0 \\ \vdots \\ \dfrac{\partial Q}{\partial \overline{\beta_m}} = \sum_{i=1}^{n} \left(y_i - \overline{\beta_0} - \overline{\beta_1} x_i - \cdots - \overline{\beta_m} x_i^m \right) x_i^m = 0 \end{cases} \tag{4}$$

The sum of residual squares Q can be used as an evaluation index to test the goodness of fit of regression equation. Generally speaking, the smaller the sum of residual squares is, the better the fitting effect is.

The quadratic fitting curves of different orders were obtained by least square fitting, as shown in Table 3. It can be seen from Table 3 that with the increase of the order of the equation, the sum of squares of residual errors gradually decreases, suggesting that the increase of order can improve the fitting degree of the regression equation to sample data. However, according to experience, over-fitting is easy to occur if the order is too high, so the order of fitting in this paper is not more than 5.

Table 3 Regression equation table

order	The regression equation	The sum of the squares of residuals Q
n=1	$y=1.036x+0.0655$	0.5295
n=2	$y=-0.333x^2+1.274x+0.05$	0.5085
n=3	$y=-0.1176x^3-0.175x^2+1.224x+0.0493$	0.5082
n=4	$y=2.5x^4-4.1455x^3+1.966x^2+0.877x+0.0592$	0.5031
n=5	$y=-1.703x^5+6.126x^4-7.167x^3+2.903x^2+0.776x+0.061$	0.5017

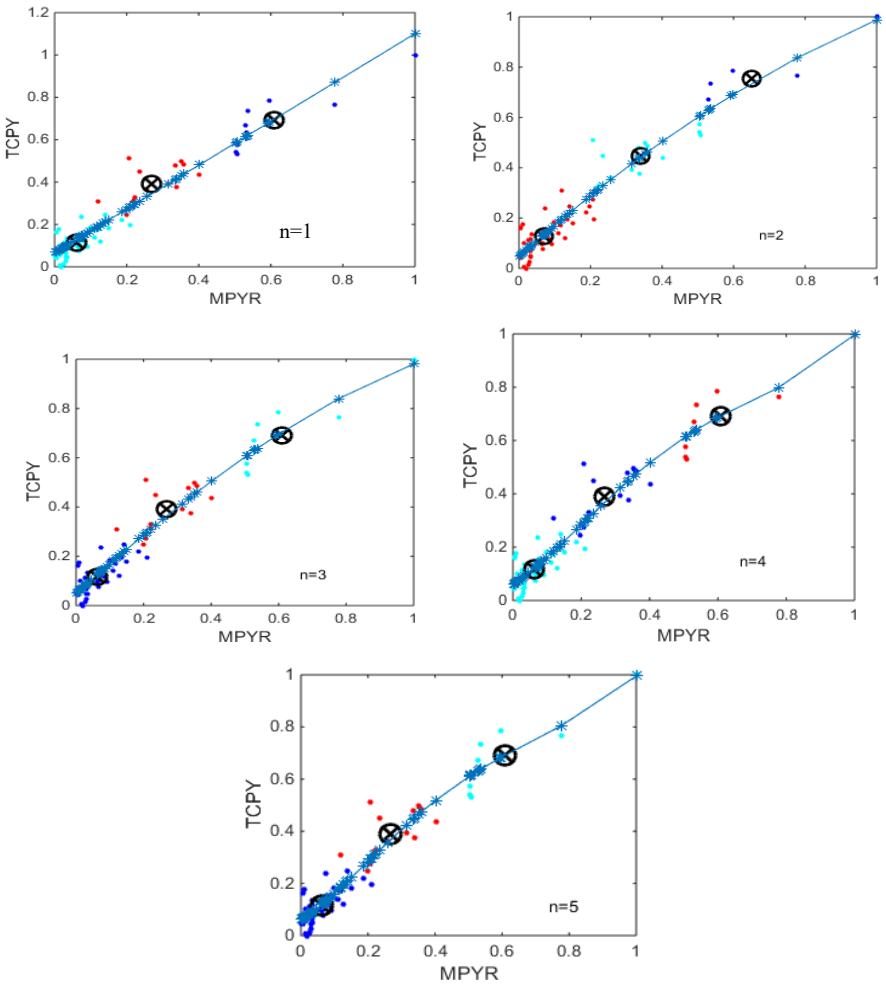


Figure 2. The curve of material person-year ratio and total cost person-year comparison.

5 MODEL VALIDATION

The regression equation is given in Table 3, which can be used to predict the total cost of the project according to the known material cost. Calculation and verification are carried out below.

Table 4, Table 5 and Figure 3 show the results of TCPY and deviation measured by the standardized MPYR. First of all, the regression formula can estimate the total project cost (excluding OPYR) according to the material cost to a certain extent.

Thirdly, although Table 3 implies that the higher the order of the regression equation is, the smaller the sum of residual squares is. The verification results show that if the order is lower, the proportion of the root mean square error in the sample mean is lower too. Therefore, the degree should be determined according to the project situation, rather than the higher order.

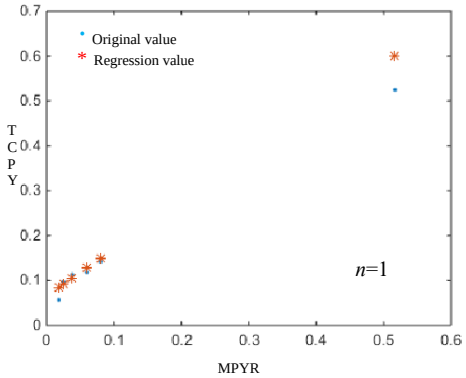
Table 4 Parameter table of each item after normalization

PROJECT	MPYR	Original TCPY	Estimated TCPY	The percentage of the deviation between the measured value and the original value in the original value while n=5.
1	0.018	0.0569	0.0762	33.90%
2	0.0245	0.0967	0.0819	15.33%
3	0.0379	0.1133	0.0945	16.62%
4	0.0596	0.1183	0.1164	1.59%
5	0.0801	0.1416	0.1386	2.11%

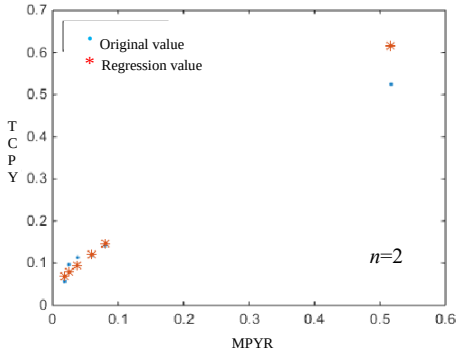
Table 5 Verification result table

Order	Residuals squared sum divided by 5	The ratio of the residuals squared divided by 6 to the sample mean
n=1	0.0334	19.05%
n=2	0.0394	22.48%
n=3	0.0402	22.93%
n=4	0.0424	24.19%
n=5	0.0418	23.84%

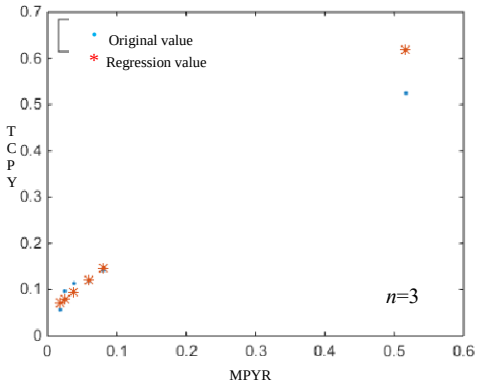
(1)


Original value
Regression value
TCPY
MPYR
n=1

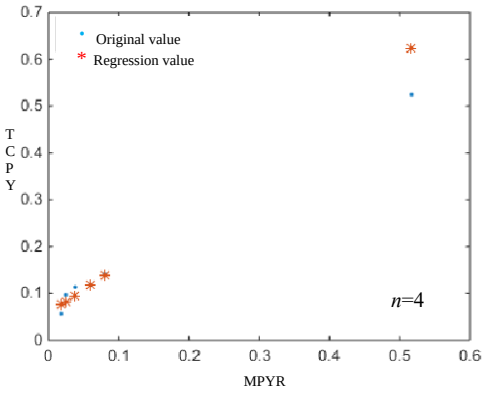
(2)


Original value
Regression value
TCPY
MPYR
n=2

(3)


Original value
Regression value
TCPY
MPYR
n=3

(4)



(5)

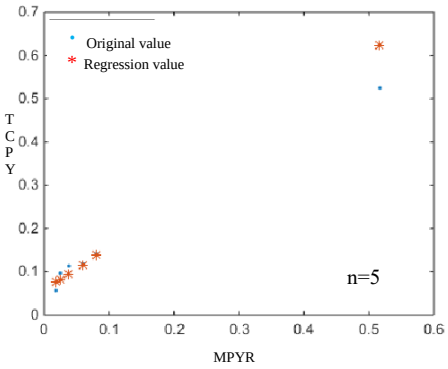


Figure 3. The curve of MPYR and TCPY comparison with different n

project cost, which requires the solution of the regression equation in Table 3. The process is similar to the above and will not be described here.

There is a large difference in the composition of development costs of different projects. The cost of general materials for the projects requiring prototype trial production is relatively high, while the cost of test and wages for partial test and research projects is relatively high. As for the cost of materials, according to reference [1], some take raw and auxiliary materials as the main body (such as steel, aluminum, titanium alloy, etc.), while others take finished products as the main body (such as computer, server, router, etc.). These differences lead to certain errors in the regression function for verifying different items. These problems should be further analyzed in the follow-up study.

The total cost parameters used in this paper exclude external cost and special cost. As external cost and special cost account for a high proportion of some projects, even more than 50%, they should be taken into account when estimating the total cost by using the equation in Table 3.

6 CONCLUSIONS

Economic data is an important support for the demonstration of equipment development costs. It is of great significance to study and master the characteristics and rules of economic data for planning the use of costs and decision-making of equipment development. In this paper, the cost data is standardized and normalized based on the analysis of the preceding equipment cost data, and then the data mining method is used to study it. Firstly, the correlation analysis method is adopted to study the correlation between the eight normalized cost data and the total cost. The cost items with weak correlation are eliminated, and only the DPYR, MPYR, TPYR and TCPY are retained as the objects of subsequent research. Then through the cluster analysis, it is found that the aggregation degree of MPYR and TCPY is the best, and the relationship is approximately linear. Then, the least square method is used to establish the 1-5 order regression model. Finally, the regression model is verified based on the actual project data. The analysis shows that the regression model can be used to estimate the equipment development cost, and the estimation accuracy is affected by the project data. In the next step, the data should be further collected and classified to lay a foundation for the in-depth analysis of the data.

REFERENCES

[1] Measures for the administration of pricing of national defense research Projects, Jan. 1995

[2] Baojun Wu, Yang Ge, Optimization method of equipment cost ratio based on combat capability, Ordnance automation, vol.34 NO.9, Sep. 2015.

[3] Meihong Liao, Ship equipment maintenance cost estimation based on data mining, Ship Science and Technology, vol.41 NO.4A, April 2015.

[4] Shouyu Song, Boqiang Liu, Research on equipment maintenance outlay demand based on data mining technology, Journal of Sichuan Ordnance Engineering, vol.31 NO.9, Sep.2010.

[5] Jianguo Pan, Lin Lin, Economic analysis of weapon equipment life cycle support system, Journal of China Electronics Research Institute, vol.10 NO.4, Aug.2015.

[6] Hongpo Wang, Yanjie Li, Xiaofeng Yu, Shaoqing Ren, Correlation analysis of eight items cost data of equipment scientific research project, The second Military Big Data Forum, Aug. 2019.

[7] Liangjun Zhang, Tan Yang, Gang Xiao, MATLAB data analysis and mining actual combat, China Machine Press, Jan. 2016.