Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/14/20230793

Maturity level determination and optimization model of D&A system based on support vector machine and multi-objective programming

Shunyu Yao

China university of geoscience, Beijing, China

1004201130@cugb.edu.cn

Abstract. In today's era of fragmented information, good data management is essential for companies. Optimizing data integration and analysis (D&A) is crucial to unleashing the full potential of data value. An evaluation was conducted on Intercontinental Freight Company's (ICM) D&A system, with three key assessments identified: importance, connectivity, and availability nature. Via multi-level fuzzy evaluation metrics, key points that affect the system's development were determined. Multiple regression lines were used to analyze relationships between people, process, and technology. Potential performance was measured via average growth in business profits, and an index of 11 variables was created to measure the system's maturity. A growth phase distribution model was developed utilizing a support vector machine and tested on a sample of 50 US ports. The optimization model, based on multi-purpose functions, maximizes efficiency and can be seen by comparing factors before and after port optimization. This experiment proves the effectiveness of the D&A system development evaluation model.

Keywords: Maturity level determination, Optimization model, Support vector machine, Multi-objective programming

1. Introduction

Assumptions were made during the model building process to better solve the complexity of the real trading and valuation process. The assumptions include: the development level of the company's D&A system is only affected by the US environment, the ICM company will strictly follow the optimization plan, and there is no technological innovation.

2. D&A system index selection and maturity measurement

2.1. Selection of performance indicators

To evaluate the three most important points, 11 different characteristics were selected including job, work and people training value, technology knowledge, difficulty, qualification, technology availability, opportunities and suitability, efficiency, safety, and ease of operation (as shown in Figure 1).

DOI: 10.54254/2755-2721/14/20230793

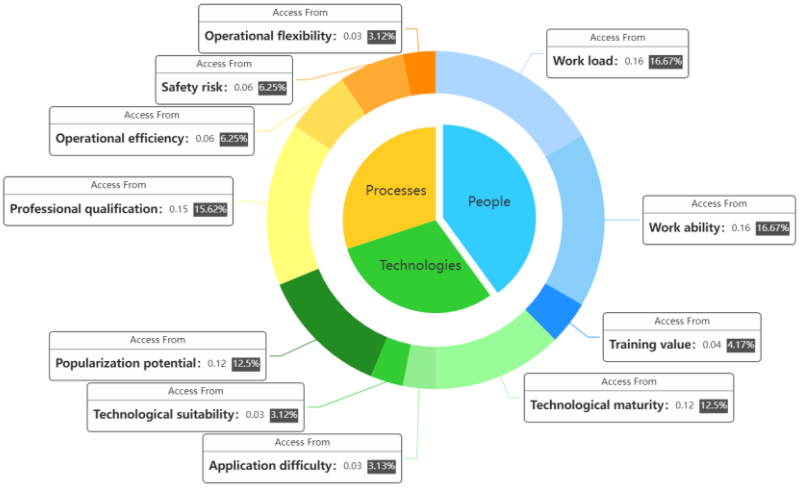

Figure 1. Characteristic variables of key factors.

People. Human, machine, material, law and environment are the five elements of production management, among which human is in the leading position and the most core position. Human is the main force of the production process, he also has the function of maintaining continuous production and improving production. Therefore, for the system's maturity assurance, talent is crucial. They should have the detailed skills to manage the system efficiently [1]. Technologies. For information technology, the most important thing is to evaluate the effectiveness of current and future technology solutions [2]. Technological Maturity, Application Complexity, Technological Appropriateness and Popularization Capability. Among them, Popularization can be an important indicator of future results, while the other three can reflect current technology [3]. Technological maturity indicates the extent to which the technology meets the desired project objectives. This is closely related to the execution process, efficiency and system reliability. So, it's a measure of technology system integrity (TS), anti-shock loading capacity (SC) and technological progress (TeP). If the technology or technology integration exists in the global market, and the combination is successful, and can play a good role in the long term, then the technology is sustainable [4]. The complexity of the application is the need to optimize the user's skills and the cost of the application technology [5]. It measures application cost (TeC), operational complexity (OX), and operational success rate (TSR). Technological development refers to the compatibility of the type of technology with the regional development goals [6], the location, financial needs, rules and regulations. For ICM companies, the right and legal (PS) mission fit, the development of technical business, law and order. The more according to the requirements of the variety, the higher the use [7]. Popularity is the possibility of sustainable use of technology in the future development process, which is important for evaluating the effectiveness of past solutions. Next [8]. Especially from the economic benefits (TeB), technology transfer (TSU), the relationship between technology and future development (CF) to consider. The more in line with the future development direction of technology, the more monopolized technology and the more income, the greater the ability to support technology [9].

DOI: 10.54254/2755-2721/14/20230793

2.1.1. Processes. Operations are critical to the entire process, which includes Productivity, Efficiency, Safety and Ease of Operation. Professional qualifications directly determine the speed of the process. Most of them come from Departmental Cooperation (DS) and Business Development (FC). Effectiveness affects maritime operations, transport, etc. Especially from the cooperation of loading and unloading lines (Co), Communication automation degree (CA), import and export trading business (SP) [10]. The security risk determines whether the process can continue. Especially from risk assessment (RM), Damage (MA), Product capability tracking (GTA), System stability (SS). Functionality is also important. Especially from data acquisition ability (DA), Time of decision (DT), Business scalability (BS). *Correlation Factors.* A D&A system is a combination of people, technology and processes. For the mature, excellent skills using data analysis tools and the ability to work effectively in the work process. The relationship between us may improve the ability to control and analyze information. Therefore, it is important to evaluate the cooperation and relationship between the main points to improve the physical ability. *Competence Factors.* D&A systems can exploit the full potential of data assets through data management and analysis. With the loading and unloading process required, the time of each connection and so on. Companies can work effectively through many factors, which is good for enhancing company strengths and gaining competitive advantage. Therefore, to see the potential of the D&A system.

2.2. Key factors determination model based on multi-level fuzzy comprehensive evaluation method
As described above, this (index) includes a variety of information. He (index) also has a taste for (good research), and participates in internal relations. (The Fuzzy Comprehensive Evaluation Method) is a good evaluation method that leads to (Fuzzy theory). It also has the advantage of providing clear conclusions, has good characteristics, and has a good way to solve complex and abstract problems. Therefore, for the difficulty (the importance of the index), but also have a good way to solve the problem important problems (Nonlinear, Fuzzy and Multivariate). Determine the key factor evaluation index system at all levels of the index element set.

$$U_i = \{U_{i1}, U_{i2}, \ldots, U_{in}\}\ (i = 1,2,3;\ x = 1,2,\ldots,n). \tag{1}$$

The set concluding the evaluation level: V=$\{v_1, v_2, \ldots, v_n\}$= “bad, poor, general, good, Excellent”. The proper way to choosing the right membership function has great effect on (The Fuzzy Comprehensive Evaluation Method). Due to the complexity of implication and accuracy, the standards vary. As a result, divide them equally to 5 intervals, each of them is name as: “Poor”, “Bad”, “General”,” Good” and “Excellent”. Such intervals are named as V, and V is equal to 5 Subset, each correspond with every basic indexes. Upon generalize these data, divide them on their values. Fuzzy Evaluation Matrix Consisting of Evaluation Indexes and Evaluation Levels:

$$R = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ r_{m1} & r_{m2} & \cdots & r_{2m} \end{pmatrix} \tag{2}$$

$0 \le r_{ij} \le 1$, i=1, 2, …, m, j=1, 2, …, n, r_{ij} represents the membership degree of the ith factor to the jth species evaluation, which can be calculated from the membership degree function. Evaluation indicator empowerment A fuzzy subset I_i is introduced from the index element set U_i as the weight or weight distribution set of the index,

$$I_i = \{a_{i1}, a_{i2}, \ldots, a_{in}\}\ (i = 1,2,3;\ x = 1,2,\ldots,n) \tag{3}$$

$a_{ix}(x = 1, 2, \ldots, n)$ represent the proportion of U_{ix} in U_i The index weights of the first, second, and third levels together make up the balance of various indexes in the evaluation process of this paper, and the weight balance of each measured at the same measurement point is 1.0.

DOI: 10.54254/2755-2721/14/20230793

The result of the fuzzy quality assessment is obtained from the fuzzy matrix and the weight is fuzzy. According to the principle of maximum membership, the final score B of each key index can be determined.

$$B_{ij} = R_{ij} \cdot I_{ij} \tag{4}$$

The fuzzy score of each key point indicator can be obtained. To some extent, it reflects the average level of D&A systems (including disused D&A systems) in US port companies in terms of people, technology and processes.

2.3. Correlation factor determination based on multiple linear regression

Based on the five-year fuzzy score of the three key points in 4.1, The least squares method was used to fit the competition scores from 2010 to 2019, and a scatter plot was also used to visualize the relationship between these elements, given that everything will be affected by the environment. It is thought that there may be a linear dependence between people and technology, and between people and processes. But there is no relationship between technology and process.

2.3.1. Model establishment. Consider a five-year fuzzy score of three factors from 2015 to 2019 for example. Consider that the impact of many evaluation models is machine and process, and the estimator is the person. a classical multiple linear regression analysis model as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{5}$$

Among them,β_0、β_1 and$\beta_2$are the parameters to be estimated. When 5 records are obtained, the model can be considered to be:

$$X = \begin{pmatrix} 1 & a_{11} & a_{1_2} \\ 1 & a_{21} & a_{22} \\ \vdots & \vdots & \vdots \\ 1 & a_{51} & \dots \end{pmatrix}, \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_5 \end{bmatrix}, S = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} \tag{6}$$

In the formula: X and Y are both the independent variable matrix, while S is the independent variable coefficient vector. Then by using the least squares method. The regression parameters with the following fraction:

$$\hat{S} = (X^T X)^{-1} X^T Y \tag{7}$$

Solve：$\beta_0 = 0.2210, \beta_1 = 0.3975, \beta_2 = -0.5023.$

2.3.2. Model testing. When it comes to the importance of designing relationships, it is necessary to pass the test before the appropriate decision is made. the F test for the regression equation and the t test for all explanatory variables.

The F test is based on the sum of squares decomposition formula, which directly tests the significance of the regression equation. The null hypothesis test is:

$$H_0: \ \beta_j = 0, j = 1,2 \tag{8}$$

After calculation, $F = 38.5772, F > F_{0.05}(2,2) = 19.0000$, the null hypothesis (9) does not hold, and the model has passed the test as a whole. The regression effect is significant, that is, there is an obvious functional relationship between the independent variable and the dependent variable. Further assumptions are made for each parameter to be estimated:

$$H_0{}^{(j)}: \ \beta_j = 0, j = 0,1,2 \tag{9}$$

DOI: 10.54254/2755-2721/14/20230793

After calculation, $t = 5.4126, t > t_{0.05}(2) = 4.303$, the null hypothesis (10) does not hold, so each variable is significantly correlated with the dependent variable. To sum up, through the F test and t test, the key factors are closely related to the process and technology. Obtaining a correlational decision model based on multiple linear regression:

$$y = 0.2210 + 0.3975 \cdot x_1 - 0.5023 \cdot x_2 \tag{10}$$

From this, the expression of the key factor is determined:

$$B = \frac{\beta_1}{\beta_1+\beta_2} x_1 + \frac{\beta_2}{\beta_1+\beta_2} x_2 \tag{11}$$

2.4. Determination of competence factor

Core competencies should directly reflect the level of development of the company's ability to create trust and attract more business to the port. For example, increasing the number of port users and increasing business. the Bureau of Transportation Statistics. Number of maritime trading nations and annual revenue for the United States, 2015-2019. After evaluation, we decided to use the average annual business revenue growth over the last five years to estimate investment performance and optimize accordingly.

$$In_i^* = \frac{In_i - max\ In}{max\ In - min\ In} \quad (i = 2015,\ldots,2019)$$

$$D = \frac{\sum_{i=2015}^{2018} \left(\frac{In_{j+1}^* - In_j^*}{In_j}\right)}{4} \quad (j = 2015,\ldots,2018) \tag{12}$$

Where In is the annual revenue from maritime trade and D is the score of competence factor.

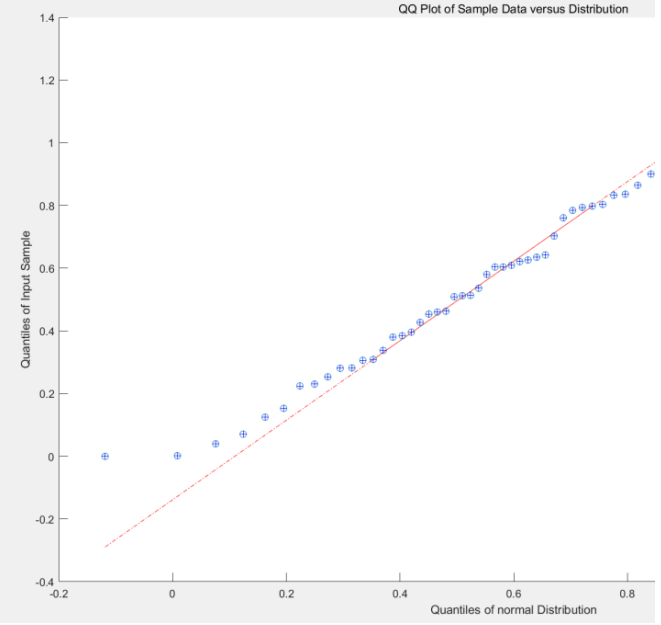

Figure 2. Quantile-quantile Plot of sample data versus distribution.

2.5. Test of the factor determination models

To verify the rationality of the model. And use the average data from 1999 to 2019 as reference to avoid the impact of the epidemic on the data. the 50 sample data needs approximately to obey any normal

DOI: 10.54254/2755-2721/14/20230793

distribution$N(\mu, \sigma^2)$.The probability that the sample data falls within $(\mu - 3\sigma, \mu + 3\sigma)$ is 99.7%. In other words, that the probability of occurrence of values other than the average 3σ is $P((x - \mu) > 3\sigma) = 0.003$, this is a very small probability event. The resulting 50 valid data sets can be used to create the next growth model. Based on the above criteria, the score of three important factors (People, technology, and process), Relationships, and Skills. To evaluate the model, the growth of the system: calculate the average value of the five indicators, divided into five equal parts of the minimum to 100% with equal parts, and then divide the feedback into five values: bad, bad, common, good, and excellent.

2.5.1. Key factors. The key factor scores of 50 seaports are shown in the figure 2. The chart analysis shows that in the middle of 30 seaports, for the same port, its three key indicators have certain regularity. The scores of each of the three key factors in the selected 50 seaports are similar. The analysis shows that 30 seaports at the average level may be in cities at similar intermediate levels, namely, similar People, Technologies, and Processes.

But some seaports in the top ten and the last ten do not conform to this rule, such as the top USLGB. It may be due to the small sample results to the large difference between seaports. A key factor, such as sophisticated technology, or a large workforce, is particularly good, potentially significantly improving the maturity of the company ' s D&A system. Similarly, a certain factor may also limit port maturity. The above is basically consistent with the reality. Therefore, from the visual results of sample key factor scores, the key factor determination model is reasonable.

2.5.2. Correlation factor. The above analysis shows that the values of the fit factors of 50 sample data approximately obey normal distribution. Therefore, the Confidence Test and Goodness-Of-Fit Test to determine the rationality of the matching factor to determine the model.

Confidence Test. For data to any normal distribution, a confidence interval with confidence level is:

$$L = \overline{X} - \frac{S}{\sqrt{n}} \cdot t_{\frac{\alpha}{2}}(n-1), R = \overline{X} + \frac{S}{\sqrt{n}} \cdot t_{\frac{\alpha}{2}}(n-1) \tag{13}$$

The bilateral confidence intervals under each rating level, and the results are shown in Table 1. The observation shows that the confidence intervals of each level are not overlapped. Therefore, it is considered that the determination model of coordination factor is reasonable and the level setting is reasonable.

Table 1. Bilateral confidence intervals at each rating level

	L	R
Excellent	0.8059	0.849
Good	0.6611	0.7354
General	0.4823	0.5453
Poor	0.2809	0.3678
Bad	0.1766	0.1921

Goodness-Of-Fit Test. QQ diagram is a good way to test the fitting degree, it is simple, intuitive, and easy to use. Therefore, the coordination factors of 50 samples, indicating the relationship between the quantile of empirical data and the quantile of the generalized distribution. These points seem to approximate a 45 degree straight line, indicating that the fitting effect is good. Therefore, it is considered that the determination model of coordination factor is reasonable. In summary, the index selection and model construction to measure the current system maturity level of ICM company. After verification, the index selection and determination are reasonable.

DOI: 10.54254/2755-2721/14/20230793

3. Suggested protocols

The 50 randomly selected port samples are mainly distributed in 27 states. According to the above optimization model, there are 22 states with improved port grade. Among them, Indiana and North Carolina increased by two grades.

Based on the model before and after optimization, for different levels, the mean and maturity scores of the five factors. After the system optimization, the maturity scores of each grade were significantly improved, and the comprehensive level of the five factors was more balanced. Therefore, through model optimization, the level of each factor in most seaports is significantly enhanced. According to the factor definition, the port has more suitable personnel, technology and process to manage the operation data, and the cooperation between them is enhanced, which is beneficial to the improvement of the company's competitive advantage. To sum up, it is necessary for companies to sign agreements in order to improve the effectiveness of the system.

ICM should trust the scientificity of our model and strictly follow the recommendations of the system maturity optimization model to improve implementation, even if it shows that a certain factor should decline. According to the coordination factor, there is a correlation between the factors. A strong ability in one aspect may even limit the development of other aspects. ICM company should pay attention to all aspects of balanced development. The best case at each level, five factor levels are fairly balanced. In the implementation of the optimization plan, the company can also calculate the mean and variance between the factors to measure the implementation of the plan in time. On this basis, the stability coefficient, that is, the development of each factor excluding its own environmental conditions. It can more intuitively reflect the development of the company since the implementation plan. It is defined as formula (20):

$$\omega_i = \frac{x_i - L_{ij}}{R_{ij} - L_{ij}} \quad (i = 1,2, \ \ldots,5; j = 1,2,\ldots,5) \tag{14}$$

Where $i$means the system level of bad, poor, general, good and Excellent. j means the five factors of people, technologies, processes, correlation, competence.

4. Model extension and discussion

For different scale seaports, their data distribution range is more extensive. To improve the applicability of the model, the classification of the system maturity rating model and set more levels to be widely used in larger or smaller sea seaports. The determination of the classification number is recommended by 1 / 10 of the sample number. In order to verify whether the above model can be applied to the wider distribution of sample data, from the United States to test the model. Ten levels are set for 100 port samples and 15 levels are set for 150 port samples. The classification effect of samples and finds that 100 and 150 sample data correspond to the same distribution rule as 50 sample data. Therefore, it is considered that our DA system evaluation and optimization model is suitable for wider data distribution, that is, according to the above model adjustment method, it can be flexibly applied to larger or smaller sea seaports.

According to the principle of system dynamics, the system is composed of structures, which determine the function of the system. Port and truck transportation belong to passenger transport system. The structure is similar, so the system functions, use the same evaluation model as the seaport to analyze and draw. It can be found that before and after optimization, the level also increased. However, different from the port company, the level of truck companies in inland states is improved, while the port is the state and city level of coastal boundary, which is in line with the distribution and business facts of the two companies, and the model is reasonable. Therefore, in the analysis of the selected indicators, the industry characteristics should be considered on the basis of the port indicators. So other related industries, through the improvement of characteristic indicators, our model can also be used to measure their maturity.

DOI: 10.54254/2755-2721/14/20230793

5. Sensitivity analysis

Combined with the correlation analysis of question 4 in Part 6, the scale range of sample data is changed, and the rating data approximately maintain the normal distribution with good reliability. This conforms to the general law of numbers in real life. That is, the system maturity level evaluation has low sensitivity to the sample range. Therefore, although only 50 samples were selected as the basis for establishing maturity level evaluation, the analysis of D&A system Maturity Evaluation model results was reliable except for the influence of rare circumstances.

For the SVM Vector product model, only use 50 port data for experiments, while only using the sharding to 0.7. In order to explore whether the system maturity score is sensitive to the number of sample data and data segmentation. At the same time, the classification prediction of vector machine is carried out with 0.1 ~ 0.9 data. The analysis shows that the accuracy of three groups of sample data for different data cutting is basically the same. Hence, the model has nothing to do with the randomness of the sample. The accuracy of each group of sample data varies significantly with the proportion of training data, so the model is sensitive to the proportion of training data. The accuracy of the three models is the highest when the training and test data ratio is 7 : 3, so the SVM model is the best when the data cutting is 0.7.

Three representative data samples to solve and demonstrate the D&A system Optimization Model. At the same time, with the average growth of the model in the last 5 years, the relationship between the situation and the change rate after development is recommended. However, in theory, changing the annual growth rate should not affect the distribution of each parameter vector. Thus, the relationship between the variance of each growth index and the average annual growth rate. When the annual growth is more than 3%, the difference between each index varies continuously with the growth rate. Abnormal data is not important if the value is less than 3%.

6. Conclusion

Our D&A system model has broad capabilities for index selection and maturity measurement. After defining development indicators. The general D&A system maturity level division model is solved by completing the maturity level distribution boundaries. Our optimized model improved factor levels and construction balance. However, the model does not consider some specific applications and requires further research. Additionally, some specific information cannot be found, requiring assumptions for problem-solving. More information can improve our model's results. The model is suitable for different data and industries, and large samples and files offer advantages.

References

[1] Angulo, C., & Català, A. (2000, May). K-SVCR. A multi-class support vector machine. In European Conference on Machine Learning (pp. 31-38). Springer, Berlin, Heidelberg.

[2] Zomorodian, M., Lai, S. H., Homayounfar, M., Ibrahim, S., & Pender, G. (2017). Development and application of coupled system dynamics and game theory: A dynamic water conflict resolution method. PLoS One,12(12), e0188489.

[3] Xiao, Y., Zhang, Y., Sun, Y., & Qian, J. (2019, October). Multi-UAV Formation Transformation Based on Improved Heuristically-Accelerated Reinforcement Learning. In 2019 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)(pp. 341-347). IEEE.

[4] Mu, C., Huang, H., & Tian, S. (2005, December). Intrusion detection alert verification based on multi-level fuzzy comprehensive evaluation. In International Conference on Computational and Information Science (pp. 9-16). Springer, Berlin, Heidelberg.

[5] Tranmer, M., & Elliot, M. (2008). Multiple linear regression. The Cathie Marsh Centre for Census and Survey Research (CCSR),5(5), 1-5.

[6] Mayor B. Growth patterns in mature desalination technologies and analogies with the energy field[J]. Desalination, 2019, 457: 75-84.

[7] Massoud M A, Tarhini A, Nasr J A. Decentralized approaches to wastewater treatment and

DOI: 10.54254/2755-2721/14/20230793

management: applicability in developing countries[J]. Journal of environmental management, 2009, 90(1): 652-659.

[8] Berler A, Pavlopoulos S, Koutsouris D. Using key performance indicators as knowledge-management tools at a regional health-care authority level[J]. IEEE transactions on information technology in biomedicine, 2005, 9(2): 184-192.

[9] Sunderland B, Burrows S, Joyce A, et al. Rural pharmacy not delivering on its health promotion potential[J]. Australian Journal of Rural Health, 2006, 14(3): 116-119.

[10] POCess A U C U B U S. NEW PRODUCT ANNOUNCEMENTS[J]. 2007.