Problem Chosen	2022	Team Control Number
D	MCM/ICM Summary Sheet	**2209617**

The Evaluation Model of Data and Analysis System

Summary

In the era of big data, many companies view data as a strategic asset, and proper management of this precious resource can bring competitive advantage. As a result, many companies need to build an integrated data and analysis (DA) system to ensure they have the right people, technology and processes to manage, manipulate, use and secure these resources.

We were hired by ICM Corporation to develop a set of reliable metrics for the company based on the information it provided, and to evaluate the maturity level of its Data and Analysis System. We scientifically demonstrate the effectiveness of the model and provide guiding suggestions for the next step of development. Finally, we study the generalization ability of the model across different scales of seaports and different domain industries.

For **task 1**, we establish an **evaluation system with a multi-layer structure**. First of all, according to the consideration of the executives of ICM Corporation, we have formulated a set of corresponding evaluation indicators to form an **index system**. The calculation containing indirect indicators adopts a variety of algorithms, including **Principal Components Analysis(PCA)**, **expert scoring** and **logistic regression**. Next, we use **entropy weight method (EWM)** to calculate the weight of each indicator, and then calculate the value of key indicators (People Index, Technology Index and Process Index) . Next, through the **group decision method(GDM)**, we use the discriminant matrix to synthesize the scores of five experts to calculate the system maturity (Maturity Index). To measure the maturity level of the system, we provide standard metrics obtained from simulated data with **K-means Clustering Algorithm**.

For **task 2**, we demonstrate in detail how our model can be used to guide the evolution of the Data and Analysis System. We not only **give operation methods**, but also carry out **case analysis** of two groups of systems. We respectively put forward suggestions for the next development of the system whose talents and technology are not up to standard.

For **task 3**, we propose a protocol for how ICM Corporation measures the maturity of the company. Since the specific data of ICM Corporation cannot be obtained, we can only solve it through **simulation analysis**. We simulated the data of similar companies and adopted the above methods to provide specific guidance for the use of ICM Corporation. Meanwhile, in order to measure the effectiveness of the system, we build a relationship model of maturity and effectiveness using the **exponential fitting algorithm with clip**.

For **task 4**, we investigated **the generalization ability of the model**. We use **sensitivity analysis** to see whether the model can be adapted to different seaports or industries. Considering the strength and weakness of the proposed model, we believe that the model is reliable.

Ultimately, we wrote a **letter** to the users of ICM Corporation, in which we introduced our work and the efforts of ICM Corporation to gain users' recognition and efforts. We hope that this letter can enhance users' understanding and confidence towards ICM Corporation's data analysis system.

Key Words: Evaluation System, EWM, GDM, PCA, K-means Clustering Algorithm

Contents

1 Introduction

1.1 Background

In the era of big data, data has become one of the core assets of companies and even the most important one. However, how to extract valuable information from massive data concerns the majority companies. To develop the essential capability of properly save and process data, data and analysis (DA) system are established.

Measuring the maturity level of the D&A system can evaluate the current situation and provide further guidance to improve the effectiveness of the D&A system and benefit the company. A mature D&A system needs to rely on talented people, reliable technology and reasonable processes to coordinate operations. The three key elements of the system (people, technologies and processes) and their connections will be the key indicators to measure the D&A system.

1.2 Restatement of Problems

As a company in charge of large-scale seaport, Intercontinental Cargo Moving (ICM) Corporation needs to rely on mathematical models to evaluate its D&A system, and expect to obtain a guidance to optimize their current work and maximize the potential of their data assets, resulting in customers' trust and approval.

We, as a consulting team, are expected to accomplish the following tasks:

- to build models for ICM Corporation to measure the current D&A system based on key performance indicators of D&A people, technologies and processed,
- to provide guidance for further improvement of the D&A system aiming to maximize their potential after the evaluation,
- to propose protocols for ICM Corporation to operate to measure the effectiveness of their D&A system,
- to analyse the generalization of proposed models in different scale of seaports or varied induestries.

1.3 Analysis of the Problem

To find an efficient method to estimate the maturity level of the current D&A system, it is essential to find a proper set of representative indicators, named as **index system**. After selecting appropriate evaluation indicators and forming the index system based on mathematical methods, the performance of D&A people, technologies and processes can be obtained with the help of entropy weight method(EWM), corresponding to the three **key performance indicators** — People Index(PeoI), Technologies Index(TechI) and Processes Index(ProcI). After that, we use the group decision making method(GDM) to study three key indicators and the relationship between them to obtain the maturity index(MI)and level(ML) of the D&A system.

Based on evaluation results of the above-mentioned model, we will offer a proposal for future development of ICM Corporation. Meanwhile, we will point out the agreements that the company should abide by in this process.

Finally, we evaluate the pros and cons of the proposed model, especially the generalization ability of the model. We carried out sensitivity analysis and reliability analysis of the model, and respectively gave the performance of the model in the face of large and small ports and different industries.

2 Assumptions and Symbol Description

2.1 Assumptions

- The government's policy on seaports will not change in the short term.
- Simulation data is reliable
- Neglect the explosive changes when forecasting over the few decades.
- Score matrix given by experts is not affected by subjective factors.

2.2 Symbol Description

We define and explain the mathematical symbols used in the following text(Tab. 1).

Table 1: Symbols

Symbol	Mathematical Meaning
X_{ij}	Original data
Y_{ij}	Normalization
p_{ij}	Conditional probability function
E_j	Information entropy
W_i	Weight of each indicator
$PeoI$	People Index
$TechI$	Technologies Index
$ProcI$	Processes Index
$MatI$	D & A System Maturity Index
$\mathrm{E}^{(I)}, E^{(II)}, E^{(III)}, E^{(IV)}, E^{(V)}$	Judgment matrix given by experts
$\mathrm{W}^{(I)}, W^{(II)}, W^{(III)}, W^{(IV)}, W^{(V)}$	Eigenvectors of the Judgment matrix
λ_i	The eigenvalues of the Judgment matrix
CI	consistency index
RI	average random consistency index
CR	consistency ratio

3 Evaluation Model for D&A System

In this section, we propose an evaluation model for D&A system based on a group of indexes. According to the operation mode and work objectives of D&A system provided by ICM Corporation, we select a series of indicators to form an Index system. Then, the three key components (people, technologies and processes) are evaluated by the entropy weight method (EWM) to obtain the corresponding indexes (PeoI, TechI, ProcI). Next, we use the group decision making method (GDM) to comprehensively consider the value of PeoI, TechI, PeocI and their connection to gain the value of D&A System Maturity Index (MI).

We obtain the original data of several companies through data simulation, and evaluate the Maturity Index of their D&A System based on the model mentioned above. Then we perform clustering analysis through the K-means clustering algorithm to quantify the level of maturity(ML).

The structure of Evaluation Model for D&A System is demonstrated in Fig.1.

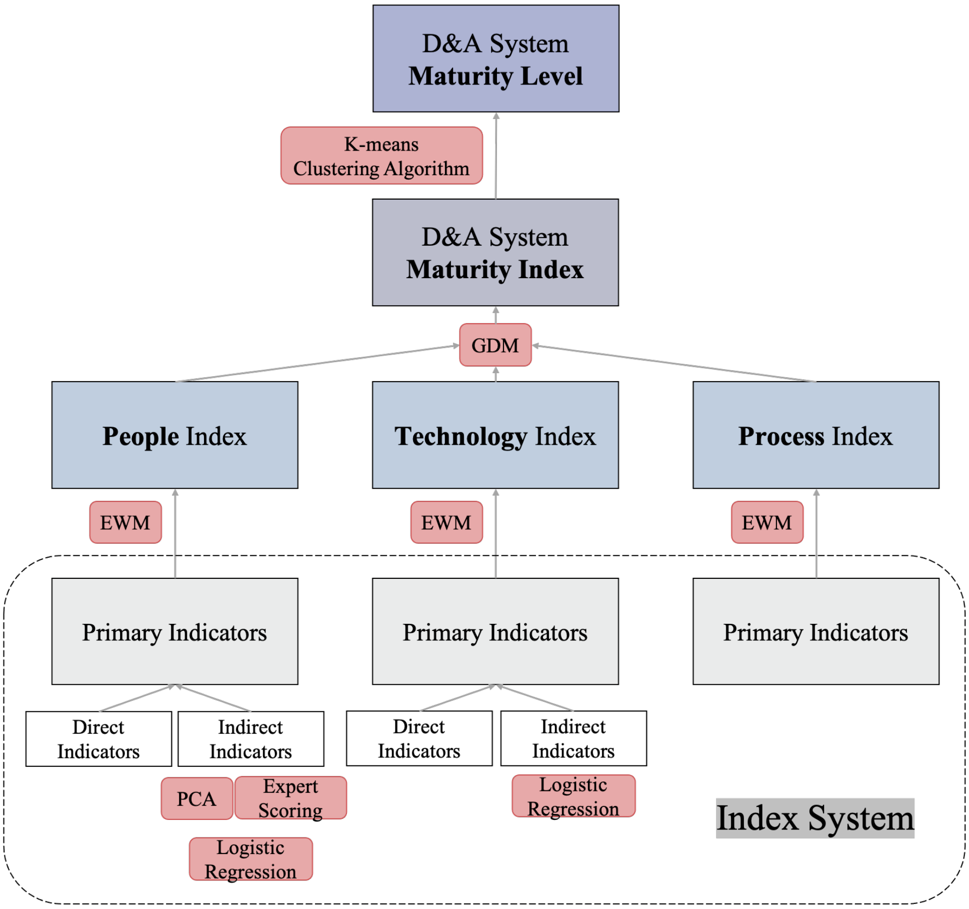


Figure 1: The structure of Evaluation Model

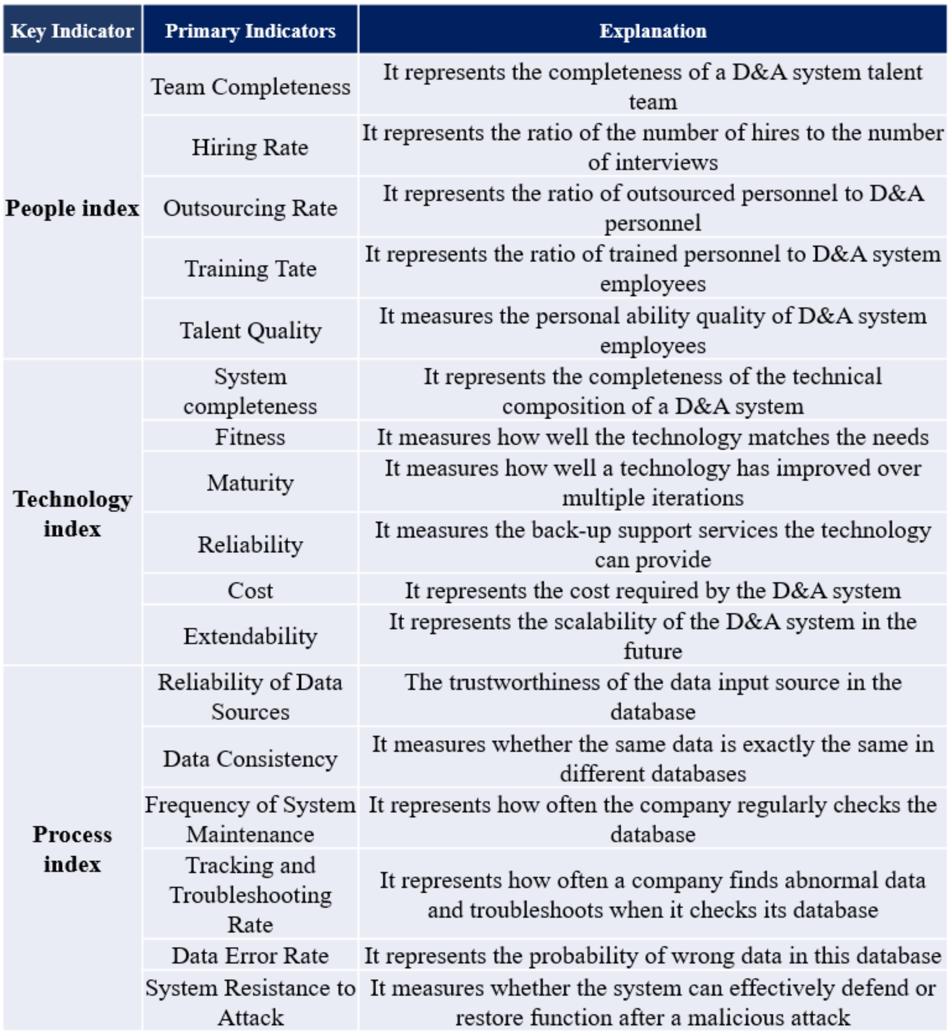
3.1 Index System

3.1.1 Index Selecting

For a mature D&A System, talented people, advanced technologies and efficient processes are required and all three aspects need to be effective and coordinated. To measure the three key elements in the D&A System, we select initial indicators for each aspect. In this section, we will introduce the corresponding metrics, the principle of choice and reasons. Primary indicators are shown as in Fig.2. Among them, indirect indicators are marked with an underline, and the detailed calculation process will be introduced in 3.1.2.

- **Primary indicators for People**

  In response to the concern of the HR team of ICM Corporation, we put forward five primary indicators of Team Completeness, Hiring Rate, Outsourcing Rate, Training Rate and Talent Quality to measure the employee quality and talent reserve of D&A system. Among them, Team Completeness is used to evaluate the talent pool of the D&A team



Key Indicator	Primary Indicators	Explanation
People index	Team Completeness	It represents the completeness of a D&A system talent team
	Hiring Rate	It represents the ratio of the number of hires to the number of interviews
	Outsourcing Rate	It represents the ratio of outsourced personnel to D&A personnel
	Training Tate	It represents the ratio of trained personnel to D&A system employees
	Talent Quality	It measures the personal ability quality of D&A system employees
Technology index	System completeness	It represents the completeness of the technical composition of a D&A system
	Fitness	It measures how well the technology matches the needs
	Maturity	It measures how well a technology has improved over multiple iterations
	Reliability	It measures the back-up support services the technology can provide
	Cost	It represents the cost required by the D&A system
	Extendability	It represents the scalability of the D&A system in the future
Process index	Reliability of Data Sources	The trustworthiness of the data input source in the database
	Data Consistency	It measures whether the same data is exactly the same in different databases
	Frequency of System Maintenance	It represents how often the company regularly checks the database
	Tracking and Troubleshooting Rate	It represents how often a company finds abnormal data and troubleshoots when it checks its database
	Data Error Rate	It represents the probability of wrong data in this database
	System Resistance to Attack	It measures whether the system can effectively defend or restore function after a malicious attack

Figure 2: The structure of Evaluation Model

which is relied on the structure of data analysis team proposed by Harvard Business School[6] for measurement. In addition, we refer to the talents evaluation system in smart manufacturing enterprise proposed in [10] to evaluate the quality of talents and obtain the evaluation indicators of Talent Quality. The calculation method of these two indicators will be introduced in 3.1.2.

- **Primary indicators for Technology**

  In response to the requirements proposed by the IT department of ICM Corporation, we have selected a set of indicators to measure the level of science and technology, which guides the D&A system how to choose the most appropriate technical solution. Among them, System completeness is the value obtained by measuring the 12 features mentioned by George Lawton in [4], and the calculation method will be introduced in 3.1.2. At the same time, based on the conclusions drawn by Manuel Vellon in [8], we also put factors such as fitness, maturity, reliability, costs and long-life/extendable into consideration.

- **Primary indicators for Process**

  For a D&A System, the security and consistency of data throughout the lifecycle is particularly important[1]. Through Brief Overview, we found that The Information

Security Officer (ISO) relies on a data governance program for data protection, management, and supervision. Therefore, in order to test the ability of D&A System to supervise functioning process, we choose to assess the ability of this data governance program. According to the function of this program, we propose corresponding evaluation indicators.

- For function of monitoring data sources, we propose Reliability of Data Sources.
- For function of tracking and approving access and changes to data, we propose Frequency of System Maintenance, Tracking and Troubleshooting Rate, Data Error Rate.
- For function of providing metadata consistency, we propose Data Consistency.
- To evaluate the safety of the program, we propose System Resistance to Attack.

3.1.2 Indirect Index calculation

In this section, we will introduce in detail how the indirect indicators that require secondary calculation (Talent Quality, Team Completeness and System Completeness) are calculated.

Talent Quality based on PCA and expert scoring

45 initial indicators are given firstly to measure Talent Quality from [10]. In order to select the important elements, the method of PCA[3] (Principal Component Analysis) is adopted. Eventually, we obtained 10 indicators with high rating and used them as an important measure of Talent Quality. The program of PCA can be found in Appendices.

We calculate the weight of each indicators mentioned above through expert scoring. We obtain expert scores on the importance of the indicators through survey materials and consulting experts. Given the result, we use multiple linear regression to obtain the indicator value of talent quality.

The corresponding expert scores and weight values of the ten initial indicators are shown in Tab. 2.

Table 2: The first ten initial indicators of Talent Quality

Indicators	Expert score	Weight
Data Sensitivity	0.9	0.1385
Data MiningAbility	0.9	0.1385
Professional Knowledge	0.8	0.1231
Information Technology Knowledge	0.8	0.1231
System Thinking	0.7	0.1077
Team Spirit	0.7	0.1077
Information Calling Ability	0.5	0.0769
Internet Thinking	0.5	0.0769
Stamina and Perseverance	0.5	0.0769
Information Integration Ability	0.2	0.0308

Team Completeness based on logistic regression

Based on [6], we establish the complete structure of the D&A talent team: Data Scientist, Data Engineer, Data Analyst, Data Manager, Data Director, and Chief Data Officer. We use the sigmoid function to perform binary logistic regression. The corresponding indicator calculation expression is shown in Eq. 1, where x is the number of vacancies.

$$TeamCompleteness = \frac{2}{1+e^{x}} \tag{1}$$

System Completeness based on logistic regression

Similar to the method of Team Completeness calculation, we perform binary logistic regression based on the 12 features proposed in [4], and obtain the calculation expression of it as Eq. 2, where x is the number of vacancies.

$$SystemCompleteness = \frac{2}{1+e^{0.5x}} \tag{2}$$

3.2 Weighting Model Based on Entropy Weight Method

In this section, we will introduce the evaluation model for obtaining the performance of the key components (PeoI, TechI and ProcI) from primary indicators using the entropy weight method(EWM).

3.2.1 Model Building

The Entropy[11] Method comes from thermodynamics and was later introduced into other disciplines. It has shown good ability in dealing with engineering technology and many other issues[9]. EWM has shown its strength in a wide range of system evaluation problems, so it is reliable to take EWM into consideration. The detailed algorithm is shown in Fig. 3.

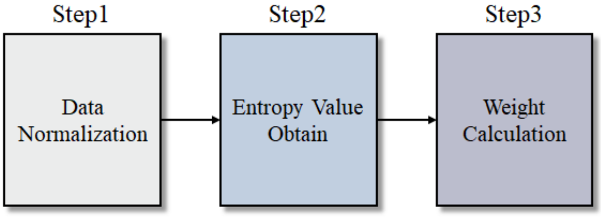


Figure 3: The structure of EWM

1. Data normalization (Min-Max Scaling):

   First, Min-Max Scaling is performed after the gain of indicator data X_{ij} , since different indicators have different dimensions. Normalization makes the features of different dimensions in the same numerical order and reduces the influence of large variance, making the model more accurate. Since indicators can be divided into positive indicators [1]and negative indicators.[2] The normalization formulas can be found in Eq.(3) and Eq.(4).

$$Y_{ij} = \frac{X_{ij} - \min_i X_{ij}}{\max_i X_{ij} - \min_i X_{ij}} \tag{3}$$

[1]**positive indicators:** The larger the value, the better.

[2]**negative indicators:** The smaller the value, the better.

$$Y_{ij} = \frac{\max_i X_{ij} - X_{ij}}{\max_i X_{ij} - \min_i X_{ij}} \tag{4}$$

2. Calculate the information entropy of each indicator:

   Then, we use the normalized data set $\{Y_{i,j}\}$ according to each indicator, which is determined by the formula 5.

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

   We can therefore calculate the probability function set $\{p_{ij}\}$ of this indicator,[2] and then calculate the information entropy of each indicator according to its definition in information theory, indicated by the formula 6.

$$E_j = -\frac{1}{lnn} \sum_{i=1}^{n} p_{ij} lnp_{ij} \tag{6}$$

3. Determine the weight of each indicator[7]:

   According to the calculation formula of information entropy, the information entropy of each indicator is calculated as

$$E_1, E_2, \ldots, E_k.$$

   We calculate the weight of each indicator through information entropy by the formula 7.

$$W_i = \frac{1 - E_i}{k - \sum_{i=1}^{k} E_i} (i = 1, 2, \cdots, k) \tag{7}$$

4. Calculate the value of primary indicators through the weight of each indicator:

   Next, we use the weight of each primary indicators to calculate the value of PeoI, TechI and ProcI, as shown in Eq. 8.

$$Index = \sum_{i=1}^{n} W_i \cdot p_{ij}, \quad i = 1, 2, \cdots, n \tag{8}$$

3.2.2 Result Analysis

Due to the protection of the confidentiality of ICM Corporation, we do not have permission to obtain real data. Therefore, to use our model, we must first build a set of simulated data based on information provided by ICM Corporation. We ran the EWM algorithm on the simulated data to obtain the weights of each primary indicator. Because of the assumption, we believe that this result can guide the systematic evaluation of ICM Corporation to a certain extent.

The weights of each indicator we calculated are shown in Tab.3. We can see that in terms of measuring the quality of talents, the weights of each item are slightly different. Among them, the impact of Outsourcing Rate is significant, indicating that outsourcing data analysis to a professional team may represent a higher level of maturity of the system and be more worthy of customer trust. At the same time, the ability of employees is also essential. As for science and technology, we can see that the value of Fitness is as high as 0.28. The full potential of data analysis techniques can only be realized when the science and technology are matched to the problem at hand. Considering the problem of dealing with the process, the frequency of abnormal data (Data Error Rate) will greatly affect the score of ProcI.

Table 3: Weight of Indicators

Category	**Primary indicators**	**Weight**
People	Team completeness	0.1205
	Hiring rate	0.1950
	Outsourcing rate	0.2465
	Training rate	0.1999
	Talent Quality	0.2381
Technology	System completeness	0.2041
	Fitness	0.2757
	Maturity	0.1622
	Reliability	0.0764
	Cost	0.1563
	Extendability	0.1254
Process	Reliability of data sources	0.1279
	Data consistency	0.1031
	Frequency of system maintenance	0.1984
	Tracking and Troubleshooting Rate	0.2112
	Data Error Rate	0.2357
	System resistance to attack	0.1236

3.3 Weighting Model Based on Group Decision Making Method

In order to obtain the value of the Maturity Index, we use the consistency judgment matrix(GDM) to weigh each key indicator. The structure of the method is shown in Fig.4.

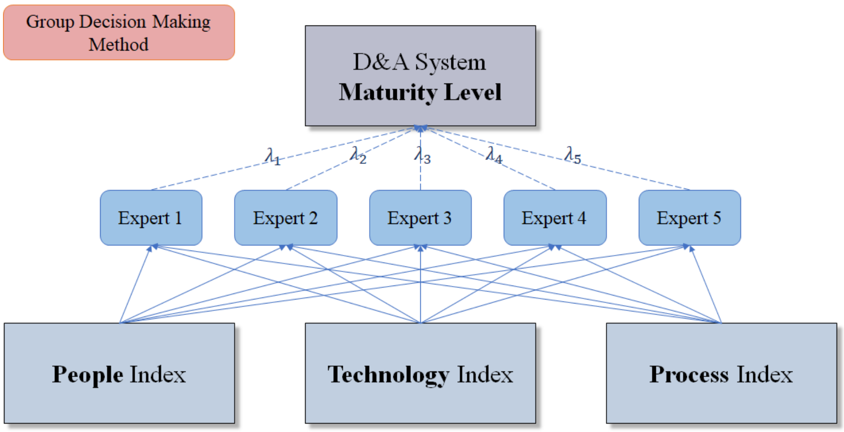


Figure 4: The structure of Group Decision Making Method

We have five experts to do their parts. For each expert, we obtain a consistent complementary judgment matrix, which is scored from the expert's experience and professional knowledge. The scaling table of judgment matrix is shown in Tab.4.

Table 4: Scaling Table of Judgment Matrix

Scalinga_{ij}	Meaning
1	i and j are equally important
3	i is slightly more important than j
5	i is more important than j
7	i is much more important than j
9	i is significantly more important than j
2,4,6,8	is the middle value of the above adjacent judgment matrix
$\frac{1}{a_{ij}}$	$\frac{1}{a_{ij}}$represents the result of comparing the j element with the i element

And the matrices following the table are given below.

$$E^{(I)} = \begin{pmatrix} 1 & \frac{1}{4} & 3 \\ 4 & 1 & 7 \\ \frac{1}{3} & \frac{1}{7} & 1 \end{pmatrix} E^{(II)} = \begin{pmatrix} 1 & \frac{1}{2} & 4 \\ 2 & 1 & 6 \\ \frac{1}{4} & \frac{1}{6} & 1 \end{pmatrix} E^{(III)} = \begin{pmatrix} 1 & \frac{1}{5} & 2 \\ 5 & 1 & 8 \\ \frac{1}{2} & \frac{1}{8} & 1 \end{pmatrix}$$

$$E^{(IV)} = \begin{pmatrix} 1 & 1 & 7 \\ 1 & 1 & 7 \\ \frac{1}{7} & \frac{1}{7} & 1 \end{pmatrix} E^{(V)} = \begin{pmatrix} 1 & \frac{1}{4} & \frac{1}{3} \\ 4 & 1 & 3 \\ 3 & \frac{1}{3} & 1 \end{pmatrix}$$

We solve the eigenvectors of the above matrix for the weights. And the group decision-making method suggests us to obtain weight from the Eq.(3.3).

$$W^* = \sum_{i=1..5} \lambda_i W^{(i)}$$

And the result of the W can be seen in the fellows.

$$W^{(I)} = (0.2109, 0.7049, 0.0842)^T, W^{(II)} = (0.3234, 0.5877, 0.0889)^T, W^{(III)} = (0.1618, 0.7510, 0.0872)^T$$

$$W^{(IV)} = (0.4667, 0.4667, 0.0666)^T, W^{(V)} = (0.1172, 0.6144, 0.2684)^T$$

Later, we check the reliability of the method to promise safely using of the proposed method by execute the following instructions of consistency test.

- calculate the consistency index value CI which should tend to zero by Eq.(9)
- calculate the consistency ratio, see Eq.(10)

$$CI = (\lambda_{max} - n)/(n - 1) \tag{9}$$

$$\begin{gathered} CR = CI/RI \\ \text{when } n = 1, .., n, \\ \text{the value of RI is given as } 0(1), 0(2), 0.52(3), 0.89(4), 1.12(5)...1.63(9), 1.64(10), \\ \text{where the numbers in brackets indicate n.} \end{gathered} \tag{10}$$

The smaller the value of CI, the better the consistency of the judgment matrix. In practical application, it is unlikely that all judgment matrices have complete consistency, and the ratio of CI to RI (average random consistency index), CR , can be used to judge whether the matrix has satisfactory consistency. If $CR < 0.10$, the judgment matrix has satisfactory consistency. The above judgment matrix is checked for consistency, and the results are shown in Tab.5.

Table 5: Judgment Matrix Consistency Test Results

Judgment Matrix	λ	CI	RI	CR
W1	3.0324	0.0162	0.52	0.03115
W2	3.0092	0.0046	0.52	0.00885
W3	3.0055	0.00275	0.52	0.00529
W4	3	0	0.52	0
W5	3.0735	0.03675	0.52	0.07067

Through the method mentioned above, the weights vector of key indicators are eventually calculated. The specific value of the weight vector is shown in Eq.(11)

$$w = [0.2929 \quad 0.5796 \quad 0.1275] \tag{11}$$

Therefore, the maturity index of the DA system can be calculated according to Eq.(12)

$$MatI = w_1 * PeoI + w_2 * TechI + w_3 * ProcI \tag{12}$$

where w_i is the weight of i th key indicator

3.4 Metric of the DA System Maturity Level

The weighting model established above can obtain the indicators of three key components (PeoI, TechI and ProcI) based on the company's underlying data, and then acquire the value of a company's DA System Maturity Index. And this section will introduce the scientific method to obtain the maturity level(ML).

We choose to use a clustering algorithm, in which the K-means[5] Clustering Algorithm is more adaptable and beautiful. Through simulation, we obtained the indicator data of nine different types of seaport companies (Tab. 6). And then, we use K-means Algorithm for cluster analysis to obtain class centres, which work as a basis for classification. We divide the DA System Maturity Level into Normal, Good and Perfect. It is worth noting that our measurement of Maturity Level does not rely on the data of the Maturity Index, but starts from three key indicators: clustering the key indicators three times, and measuring the Maturity Level based on the relationship between them.

Table 6: The result of simulated system

system index	PeoI	TechI	ProcI	system maturity level
1	0.208236066	0.273421277	0.32549437	0.260968
2	0.223996076	0.323155973	0.327655843	0.294686
3	0.184314895	0.292679262	0.386015942	0.27284
4	0.418640529	0.274986239	0.461780003	0.340879
5	0.545884401	0.419500309	0.460997093	0.461809
6	0.627048778	0.352171461	0.56200426	0.459437
7	0.711752974	0.466277884	0.634438651	0.559618
8	0.679837004	0.717203797	0.705845554	0.704811
9	0.725638581	0.762995978	0.614811072	0.73316
Weight	**0.2929**	**0.5796**	**0.1275**	

3.4.1 K-means Clustering Algorithm

The detailed algorithm is shown as follows:

Algorithm 1 K-means Clustering Algorithm

```
Input: Data set $X = x^{(1),x^2,\cdots,x^(m)}$
Output: Cluster centroids $\mu_{i=1,\cdots,k}$; Cluster assignments $c \in \mathbf{Z}$
Initialize $k$ cluster centroids $\mu_1, \cdots, \mu_k$ randomly from X;
repeat
    for $i = 1, \cdots, m$ do
        set $c^{(i)} = argmin_j \left\| x^{(i)} - \mu_j \right\|^2$;
    end for
    for $j = 1, \cdots, k$ do
        set $\mu_j = \frac{\sum_{i=1}^{m} 1\{c^{(i)}=j\}x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)}=j\}}$;
    end for
until Convergence;
return $\mu$ and $c$;
```

3.4.2 Result Analysis

In order to properly measure the Maturity Level, we first clustered the three key indicators on the simulated data to obtain the standards for measuring the level of the key indicators. We obtain the Maturity Level standard by analyzing the value and correlation of key indicators. This is because a mature system not only needs to have good performance in all aspects, but also needs to be consistent. The results obtained using the K-means clustering Algorithm are shown in Fig.5

We divide it into three indicators, Normal, Good and Excellent, where Normal means that the system is in its infancy and not yet mature. Good says that the system has certain development prospects, and Excellent means that the system has matured and can undertake reliable data analysis work.

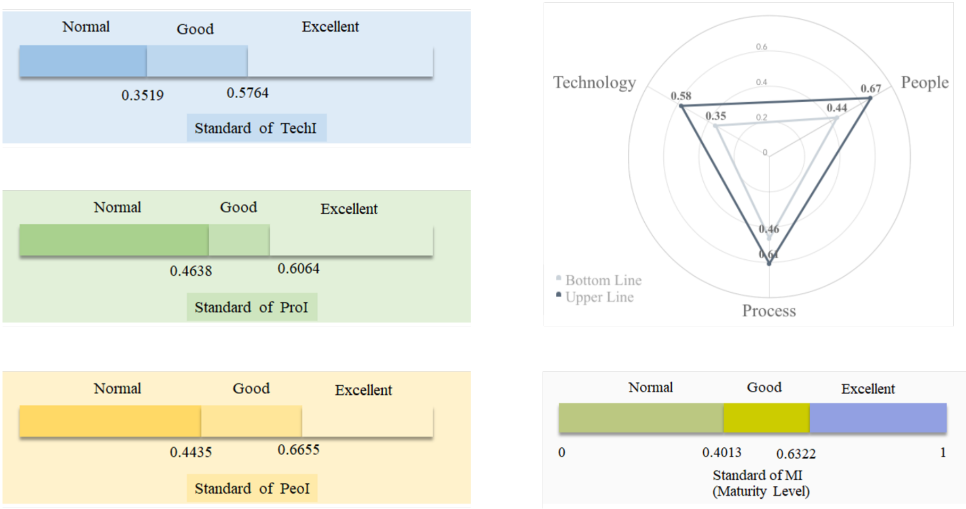


Figure 5: The standard line of Maturity Level

4 The Suggestion for D&A System Future Development

In this section, we will describe how the company uses the model established in 3 to assess the maturity of the D&A System, and how to apply the results of the model to require guidance for further development. We will first give the specific operation method, and then will give two sets of simulated data as a case for analysis, and give the corresponding development planning reference.

4.1 Suggestion

After using the Evaluation Model in 3 for evaluation, we can obtain three key elements of the company's D&A System: PeopI, TechI, ProcI. Further, we can build a radar map of them for intuitive analysis and judgment.

According to 5, we can know that the maturity measurement of D&A System is closely related to the performance of PeopI, TechI, and ProcI. For a mature system, we need not only the three indicators to have good performance, but also the coordination and matching of them. In order to give opinions on the next development of D&A System, we need to assess the indicators of PeopI, TechI, and ProcI and the consistency of them.

From the radar chart, we can determine the worst-performing indicator and calculate the optimal solution based on the stability of the triangle.

Table 7: the optimization order of primary indicators

People Index	**Rank**	**Technology Index**	**Rank**	**Process Index**	**Rank**
Talent Quality	1	Fitness	1	Data Error Rate	1
Training Rate	2	Team Completeness	2	Tracking and Troubleshooting Rate	2
Hiring Rate	3	Maturity	3	Frequency of System Maintenance	3
Team Completeness	4	Cost	4	Reliability of Data Sources	4
		Extendability	5	System Resistance to Attack	5
		Reliability	6	Data Consistency	6

Determining the direction that needs to be further improved, we recommend that the company optimize the sub-factors according to the weight obtained by 3.2.2 from large to small, that is, optimize in the order of the indicators shown in Tab.7.

4.2 Case Analysis

In this section, we obtain the initial indicator data of the two D&A Systems through data simulation. We will use the evaluation model established in 3 and the analysis method mentioned in 4.1 to simulate and analyze these two systems, obtain the maturity performance of them, and make further development suggestions.

First, we gain the values of the primary indicators that affect the D&A system from the data simulating program, see the appendix. After that, according to the evaluation model established in 3, we calculate and obtain PeopI, TechI and ProcI of the two systems respectively, and the Maturity Index of D&A Systen, shown in Tab.8.

Table 8: The Result of example system

System	Key indicators	Score	Maturity Index
1	People Index	0.1838	0.5443
	Technology Index	0.7078	
	Process Index	0.6301	
2	People Index	0.5756	0.4138
	Technology Index	0.2896	
	Process Index	0.6065	

We draw a radar chart(Fig.6) from the data in Tab.8. From the figure, we can see that the performance of system 1 in terms of technology, and system 2 in terms of talent performance is not been satisfactory. Therefore, the suggestions we gave for the future development of System 1 and System 2 are from these two aspects.

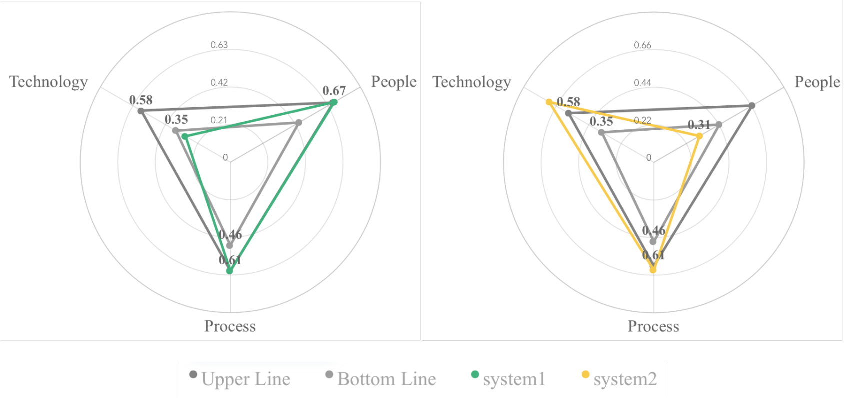


Figure 6: The Radar Chart of example systems

4.2.1 System with poor technology

function simulation and sensitivity analysis, we obtained the impact of changes in different indicators on the results. For System 1, we can see that its indicators in terms of technology are slightly lower. After analysis, we judged the change in System Completeness and Fitness will be more sensitive to the results of the system. From the prediction results (Fig.8), we give the following suggestions:

- Vigorously improve the rate of System Completeness. Following solutions are recommended: choose a DA system with multi-function and multi-features, choose a product with mixed functions or a combination of multiple products.
- According to the problem and intended purpose, choose the appropriate scientific technology, when the fitness is higher than 0.8, the data mining system will have a good performance.

4.2.2 System with less talents

For systems like system 2, where talents are not qualified, we also use the 4.2.2 method to predict and analyze. Based on the results obtained (Fig.7), we can offer guidance as follows:

- Improve the completion rate of the team, complete the staffing, and equip a data analysis team with qualified stuff in each position.
- According to the talent evaluation criteria, recruit more talented people, including but not limited to being sensitive to data and having excellent data mining capabilities.
- Increase the training rate. If money is not a limitation, it will have a better performance if the rate is increased to 0.9.

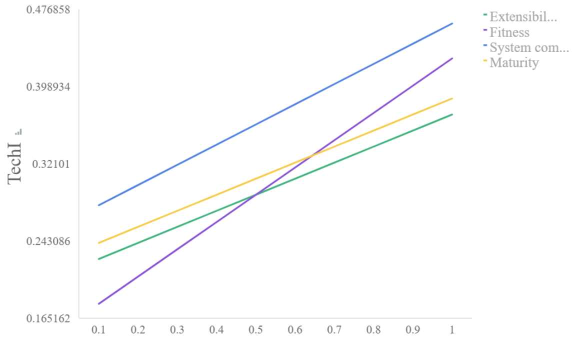


Figure 7: The result of sensitivity on system with poor technology

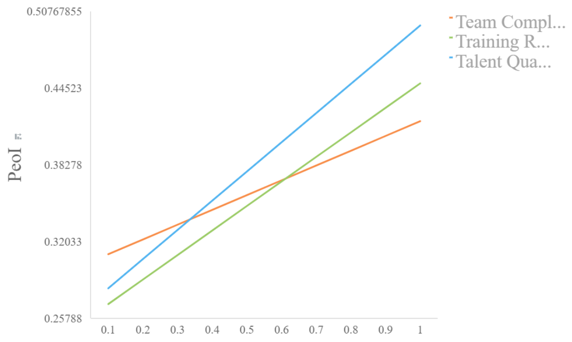


Figure 8: The result of sensitivity on system with less talents

5 Protocols to measure the effectiveness

In this section, we will address the third task: proposing a set of protocols that ICM Corporation should implement to measure the effectiveness of a D&A system. We relate the effectiveness of a D&A system to its maturity. We believe that the D&A system will be effective only after the maturity of the system reaches a certain standard, and there is a functional relationship between the effectiveness of the D&A system and its maturity in a certain stage. Therefore, we propose the exponential fitting algorithm with clip as a metric for ICM Corporation to measure the effectiveness of its D&A system.

In , we use qualitative descriptions such as normal, good and excellent to measure maturity, called the standard line. Our proposed protocol also uses the uniform partition to clip. The specific provisions are as follows:

- D&A System whose maturity level is normal, we think its maturity level is too low, it is called an invalid system, and its effectiveness is 0.
- D&A System with good maturity level, its effectiveness and maturity are exponentially related.
- D&A System whose maturity level is excellent, it is a fully effective system, and its effectiveness is 1.

Based on the protocol, we build a model, shown in Eq.(13), where f is effectiveness and ρ is maturity index(MI).

$$f = \begin{cases} 1 & \rho > s_2 \\ f(\rho) & s_1 \leq \rho \leq s_2 \\ 0 & \rho < s_1 \end{cases} \tag{13}$$

$$f(\rho) = ae^{\rho}$$

And

$$s_1 = 0.340879$$
$$s_2 = 0.704811$$
$$a = [\tfrac{1}{e}]^{s_2}$$

Eq.(13) can be written as Eq.(14).

$$f = \begin{cases} 1 & \rho > s_2 \\ (\frac{1}{e})^{s_2} \times e^{\rho} & s_1 \leq \rho \leq s_2 \\ 0 & \rho < s_1 \end{cases} \tag{14}$$

6 Model adaptation and generalization analysis

In this section, we will address task four: adapt our model to different types of seaports and companies in different fields, then study the adaptation and generalization of our model.

6.1 For the seaports in different sizes

For seaports with different specifications, we investigated and simulated the data of the data analysis systems corresponding to different size. We analyzed and calculated the maturity level based on the Evaluation Model. Result is shown in Tab.9. We comprehensively compared the relevant data, and considered the generalization ability of the model through the change of the result value. The related figure is shown in Fig.9.

We can compare and find that for large, medium and small ports, the proportion of different data have a significant impact. Among them, the changes in the weights of various indicators

Table 9: Comparison of weights between different types of seaports

category	Primary indicators	weight(O)	weight (S)	weight(M)	weight(L)
People	Team completeness	0.1205	0.1821	0.3663	0.1636
	Hiring rate	0.1950	0.1911	0.1541	0.1660
	Outsourcing rate	0.2465	0.1999	0.1788	0.3274
	Training rate	0.1999	0.1867	0.1490	0.1793
	Talent Quality	0.2381	0.2402	0.1519	0.1636
Technology	System completeness	0.2041	0.1838	0.1199	0.1172
	Fitness	0.2757	0.1484	0.1168	0.1020
	Maturity	0.1622	0.1812	0.1130	0.1017
	Reliability	0.0764	0.1691	0.2687	0.2729
	Cost	0.1563	0.1691	0.2687	0.2729
	Extendability	0.1254	0.1484	0.1139	0.1332
Process	Reliability of data sources	0.1279	0.0874	0.1265	0.1139
	Data consistency	0.1031	0.2077	0.1374	0.1044
	Frequency of system maintenance	0.1984	0.2077	0.3266	0.1114
	Tracking and Troubleshooting Rate	0.2112	0.2077	0.1236	0.2800
	Data Error Rate	0.2357	0.2077	0.1265	0.2800
	System resistance to attack	0.1236	0.0817	0.1594	0.1102

of talent are basically the same, which is in line with the actual situation: different ports need to rely on suitable talent matching teams. We also found that the neutral and large ports have little difference in TechI, which is also in line with the actual situation. The results can promise the suitability of our model in different scale of seaports.

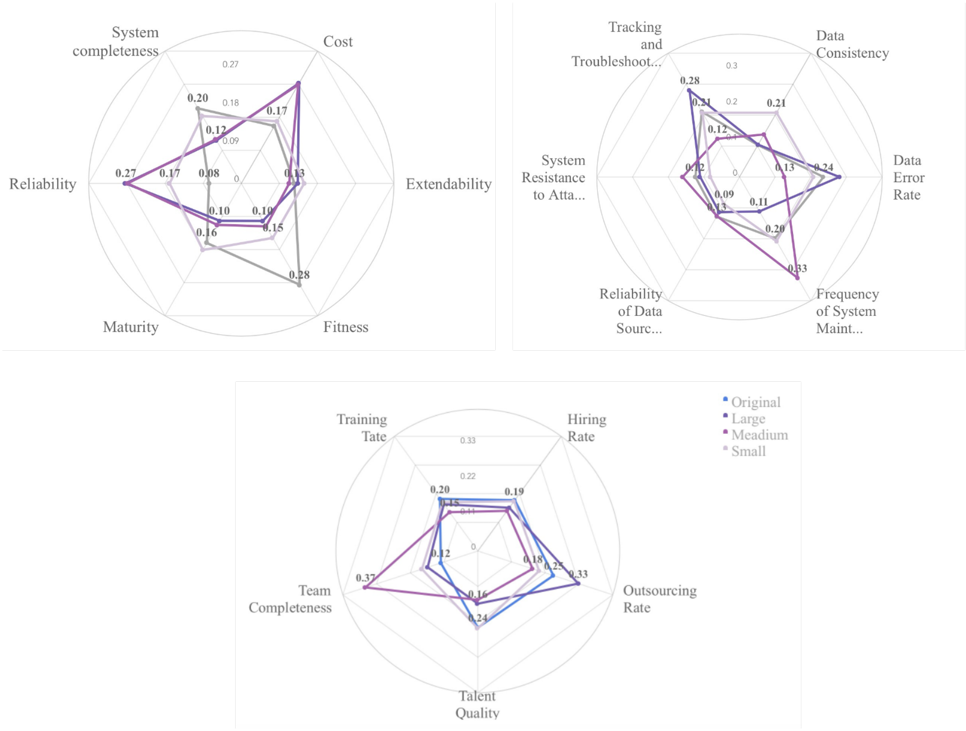


Figure 9: Weighted radar map for different types of seaports

6.2 For different industries

Companies in different fields are naturally affected by their fields in the composition of the D&A system. Therefore, the model weights obtained from the data of management seaport companies such as ICM Corporation obviously cannot be directly applied to companies in other fields. We will take trucking company as an example, recalculate and adjust the weights based on the D&A system data of them, so that the model can be properly applied to the freight company after adjustment.

According to the simulation data of the trucking company, the weights obtained by readjustment are in Tab.10. The related figure is shown in Fig.10.

Table 10: Comparison between the weights of ICM and freight companies

category	Primary indicators	weight(P)	weight (H)
People	Team completeness	0.1205	0.3387
	Hiring rate	0.1950	0.2123
	Outsourcing rate	0.2465	0.1484
	Training rate	0.1999	0.1581
	Talent Quality	0.2381	0.1425
Technology	System completeness	0.2041	0.1253
	Fitness	0.2757	0.1546
	Maturity	0.1622	0.1428
	Reliability	0.0764	0.1262
	Costs	0.1563	0.2508
	Extendability	0.1254	0.2203
Process	Reliability of data sources	0.1279	0.1756
	Data consistency	0.1031	0.1816
	Frequency of system maintenance	0.1984	0.1465
	Tracking and Troubleshooting Rate	0.2112	0.1442
	Data Error Rate	0.2357	0.1573
	System resistance to attack	0.1236	0.1849

From Tab.10 and Fig.10, we found that trucking companies should pay more attention to the completion rate and employment rate of the team, and attach great importance to the construction of the D&A system. By adjusting the weight of these indicators, our model can have more efficient and accurate evaluation for the D&A system of trucking company.

We believe that if ICM Corporation's customers adopt our model, they can accurately analyze the problems existing in their own D&A system, and make targeted improvements to the existing defects, which can effectively improve the work efficiency of them — to increase and speed up the storage of goods and the speed of transportation. Thereby that can indirectly improving the benefits of management seaport companies like ICM Corporation.

Our model can be advantageous for data-sensitive companies. Different from the general measurement of IT systems, in addition to the measurement of technology and talent, we also consider the effectiveness of data supervision. Therefore, based on the above results, we judge that our model is promising.

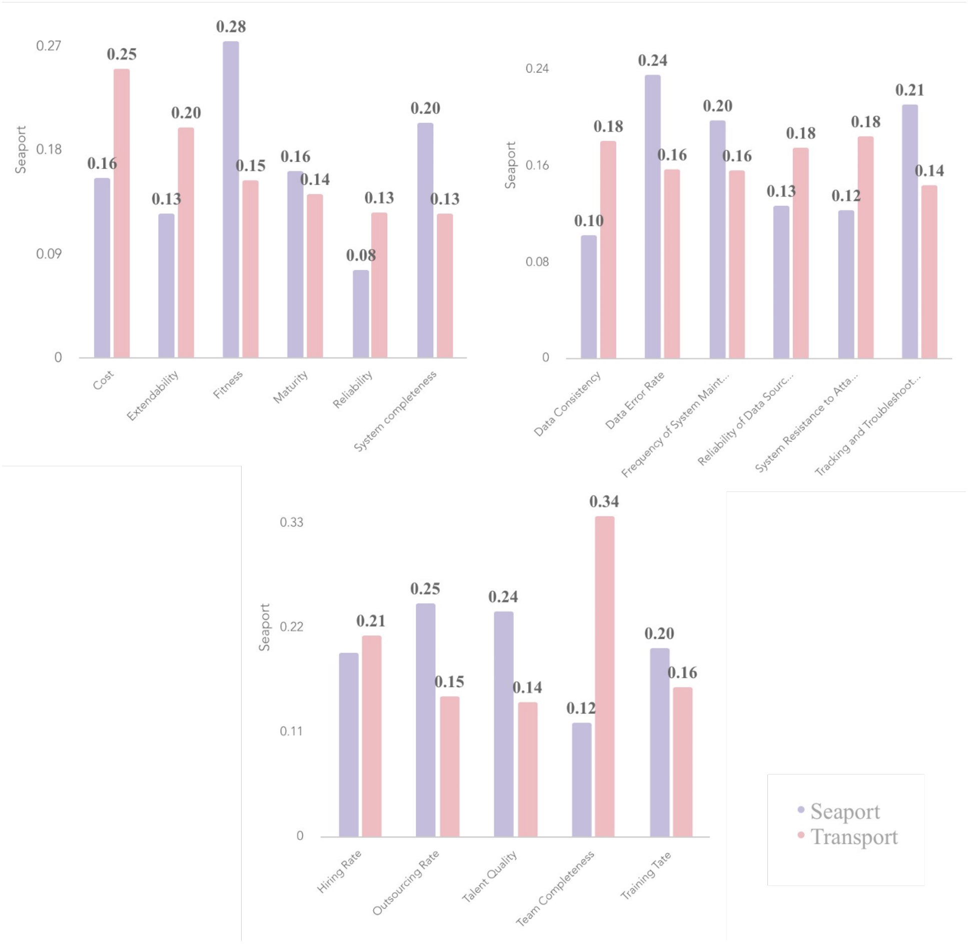


Figure 10: The weights of ICM and trucking companies

7 Strengths and Weaknesses

7.1 Strengths

- In the evaluation, grading indicators are adopted, which is highly comprehensive and the index selection is representative and can be widely applied.

- Adopt objective weighting method (entropy weight method) and the subjective (group decision making method). More effective weights are obtained.

- Using radar charts, column charts, line charts, etc.reflects the difference in weights between different seaports. The results are displayed more vividly.

7.2 Weaknesses

- Because the factors affecting the D&A system are too numerous and complicated, considering the limited time and technical environment, we only can select seventeen important factors for evaluation and testing.

- Some data which is based on simulation may bring some errors.

8 A Letter

MCM/ICM Contest

#220
9617

Consulting
Team
#2209617

Team #2209617
Problem D

Monday, 21st February 2022

A Letter to the consumers of ICM Corporation

To whom it may concern,

How are you?

We are the professionals from consulting team #2209617, who are hired by ICM Corporation to measure their Data and Analysis System and purpose suggestions on further development. We built a series of models to measure the maturity level and validity of ICM Corporation's Data and Analysis System. It is proved that our evaluation method is reliable and efficient, according to the result of reliability analysis. Understanding your concern as ICM Corporation's consumers, we are writing this letter to give a detailed introduction of our solid evaluation method. And we hope this letter can allay your worries and give you confidence in ICM Corporation's Data and Analysis System.

Relying on the information provided by ICM Corporation and the characteristics of the company's business, we have carefully selected a series of primary indicators that may affect the performance of the system, resulting in the establishment of index system. In terms of talent pool, we not only measured the integrity of the team based on the standards provided by Harvard Business School, but also measured the talent quality of the team using methods such as PCA and expert scoring. At the same time, according to the professional advice from Forbes Technology Council, we applied a mathematical method to measure the effectiveness of science and technology. We also analyzed and evaluated the flow of data in the system, using Frequency of system maintenance, Tracking and Troubleshooting Rate, System resistance to attack and other indicators to measure the security and consistency of data in D&A system.

We recommend that ICM Corporation uses EWM to measure the utility value of each indicator. Moreover, we provided a set of weighting values for their reference. Based on the method, we can obtain the performance of three key indicators (people, technology and process) of the D&A system.

Next, in order to measure the maturity rate of the system from the performance of key indicators, we adopt the method of group decision-making, referring to the opinions of five experts in the industry, and obtained the value of the maturity rate. Comparing it with the standard line assessed from K-means Clustering Algorithm, we can declare the maturity level of ICM Corporation's Data and Analysis System.

The structure of our proposed model can be found in the picture attached to this letter. After a series of generalization analysis and practical application, we can safely announce that the model is valid and the performance is promising. The detailed model-building process and test result can be found in our paper handed to MCM/ICM Contest.

We understand your concern towards the reliability of the Data and Analysis system of ICM Corporation. We can guarantee that the model we built has its strength. Additionally, we also provide reasonable suggestion on the future development of ICM Corporation which can assure you the extendable ability of the Data and Analysis System to support new requirements over the next three, five or ten years.

We hope this brief introduction will convince you of the reliability of our evaluation system, resulting in the believe that ICM Corporation attaches great importance to the data analysis business and its willingness to continuously improve its performance. Hopefully we can help you realize that ICM Corporation is a company that cares about customer satisfaction and confidence, and they are constantly improving to achieve industry-leading levels.

Thank you for reading this letter and we wish you all the best!

Yours sincerely,
#2209617

References

[1] Alvaro A. Cárdenas, Pratyusa K. Manadhata, and Sreeranga P. Rajan. “Big Data Analytics for Security”. In: *IEEE Security Privacy* 11.6 (2013), pp. 74–76. DOI: 10.1109/MSP.2013.138.

[2] Wang Cui, Wang Xingfen, and Zhuang Wenying. “The Research on Cross-border Online Shopping Transaction Risk Based on Online Data Access”. In: *2019 IEEE International Conference on Big Data (Big Data)*. 2019, pp. 5331–5335. DOI: 10.1109/BigData47090.2019.9005727.

[3] Ian T Jolliffe and Jorge Cadima. “Principal component analysis: a review and recent developments”. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374.2065 (2016), p. 20150202.

[4] George Lawton. *Twelve must-have features for big data analytics tools*. https://searchbusinessanalytics.techtarget.com/feature/12-must-have-features-for-big-data-analytics-tools. 2021.

[5] Aristidis Likas, Nikos Vlassis, and Jakob J Verbeek. “The global k-means clustering algorithm”. In: *Pattern recognition* 36.2 (2003), pp. 451–461.

[6] Tim Stobierski. *HOW TO STRUCTURE YOUR DATA ANALYTICS TEAM*. https://online.hbs.edu/blog/post/analytics-team-structure. 2021.

[7] Di Tian. “The Strategy to Measure the Effectiveness of Smart City Growth Model”. In: *IOP Conference Series: Earth and Environmental Science* 567.1 (Sept. 2020), p. 012003. DOI: 10.1088/1755-1315/567/1/012003. URL: https://doi.org/10.1088/1755-1315/567/1/012003.

[8] Manuel Vellon. *Twelve Factors To Help You Evaluate Potential Technical Solutions*. https://www.forbes.com/sites/forbestechcouncil/2017/02/09/12-factors-to-help-you-evaluate-potential-technical-solutions/?sh=9a2f0894f665. 2017.

[9] Jun Wang Wenjin Ke. “Evaluation of Urban Higher Education Resource Carrying Capacity Based on Entropy TOPSIS Model”. In: *Economics and Management Sciences* (2020). DOI: 10.13546/j.cnki.tjyjc.2020.18.011.

[10] XING LinglingWANG XinZHAO Yawen. “Construction oninnovative talents evaluation system in intelectual manufacturing enterprises”. In: *Journal of Shenyang Normal University (Natural Science Edition)* 38.1 (2020), p. 7.

[11] Yuxin Zhu, Dazuo Tian, and Feng Yan. “Effectiveness of entropy weight method in decision-making”. In: *Mathematical Problems in Engineering* 2020 (2020).

Appendices

Here are some simulation program we used in our model as follow.

The Python program used for data processing and entropy weight calculation:

```
#In this program, we calculate the entropy weight of the
# three-level indicators according to the  three-level indicator
#  data of personnel, technology and process respectively.
import numpy as np
import pandas as pd
fp="d:/P1.xlsx"
data=pd.read_excel(fp,index_col=None,header=None)
data = (data - data.min())/(data.max() - data.min())
m,n=data.shape
k=1/np.log(m)
yij=data.sum(axis=0)
pij=data/yij
#The second step, calculate pij
test=pij*np.log(pij)
test=np.nan_to_num(test)
ej=-k*(test.sum(axis=0))
#Calculate the information entropy of each indicator
wi=(1-ej)/np.sum(1-ej)
print(wi)
```

The Python program used for K-means Clustering:

```
#In this program, we cluster companies’ D&A systems based on their
# people, technology, process scores.
from sklearn.cluster import KMeans
import numpy as np
X = np.array([[0.06099, 0.15847,0.0415], [0.06561, 0.18730,0.04178],\
[0.05400,0.16964,0.4922], [0.12262, 0.15938,0.05888], \
[0.15989, 0.24314,0.05878], [0.18366, 0.20412,0.07166], \
[0.20847,0.27025,0.08089],[0.19912,0.41569,0.09000],\
    [0.21254,0.44223,0.07839]])
kmeans = KMeans(n_clusters=3, random_state=0).fit(X)
print(kmeans.labels_)
print(kmeans.cluster_centers_)
kmeans.predict([[0.06099, 0.15847,0.0415], \
    [0.06561, 0.18730,0.04178],[0.05400,0.16964,0.4922]])
```