International Journal of Database Theory and Application
Vol.9, No.9 (2016), pp.263-274
http://dx.doi.org/10.14257/ijdta.2016.9.9.24

# Construction and Simulation on Environmental Quality Evaluation Model based on Data Mining and Correlation Analysis

Meimei Wang[1,*], Duoyong Zhang[2] and Huimei Xu[3]

[1] *College of Earth and Environment Sciences, School of geological science and mineral resources, Lanzhou University, Lanzhou 730000, China*
[2] *Science and technology department of Longdong University, Qing Yang city, Gansu province, 745000, China*
[3] *College of Earth and Environment Sciences, Lanzhou University, Lanzhou 730000, China*
**Corresponding Author: wangmm13@lzu.edu.cn*

***Abstract***

*In this paper, we conduct research on environmental quality evaluation model based on data mining and correlation analysis. Along with the application of multi-statistical analysis method, the big data analysis law by has been applied in the environmental quality evaluation. Reciprocities of this method among from many targets starts that changes into a few not related overall targets many targets and the merit lies in had considered the relevance among various targets that can maximum limit retain original information, carries on best comprehensive dimensionality reduction processing to the high dimensional data. Aside by using this feature, this paper proposes the data mining and correlation analysis based model. The basic task of the analytical grey incidence is the microscopic or macroscopic geometry of behavior based factor sequence is close, to analyze and contribution degree of influence or the factor between determination factors to main behavior, but the gray incidence space carries on the foundation of analytical grey incidence. We implement the model on the air and water quality evaluation which are assisted with the neural network and gray analysis. The experimental result reflects the effectiveness of our model, it can evaluate the environmental quality effectively.*

***Keywords:*** *Environmental Quality; Data Mining; Correlation Analysis; Evaluation Model; Mathematical Optimization*

## 1. Introduction

The fundamental goal of social development is the survival and development of person needs and reality satisfies the degree, the enhancement of the people quality of life is the ultimate objective that the social development pursues and highest principle. Quality of the life index system is divided into two categories: objective conditions and subjective feelings of indicators. The objective conditions include population fertility and mortality, household income and the consumption levels, types and quality of products, employment conditions, living conditions, environmental conditions, education, health facilities and the conditions, community groups and participation rates, social security or social security and so on. Through the comparative analysis of these objective comprehensive indicators, we can weigh the degree of social change.

The subjective feeling index mainly determines people's life satisfaction and happiness determined by some demographic conditions, interpersonal relationships, social structure, psychological conditions and some other factors satisfaction is usually measured by the satisfaction of the overall life satisfaction and specific aspects of the satisfaction of two. In recent years, China's rapid economic development caused widespread environmental

ISSN: 2005-4270 IJDTA

problems, developed areas due to the large population, a variety of industries and man-made activities and environment is particularly serious. Urban environmental problems not only become the "bottleneck" and obstacle of the regional economic development, but also seriously affect the process of China's overall economic and the social development. Under this condition, in this paper, we conduct research on the environmental quality evaluation model based on data mining and correlation analysis. The rest of the paper is organized as follows. In the Section 2, we theoretically analyze the data mining methods and data classification methodologies. In the Section 3, we propose the novel correlation analysis model to serve as the basis for our systematic research. In the Section 4, we propose novel environmental quality evaluation model and conduct numerical analysis. In the Section 5, we summarize the work and give the conclusion.

## 2. The Data Analysis Paradigms

### 2.1. The Data Classification Methodology

The classification is in pattern recognition, machine learning as well as data mining a basic and important issue. Therefore, some effective classification algorithms arise at the historic moment. Due to the complexity of dealing with the incomplete data, the previous classifiers are mostly for the complete data. However, for a variety of reasons, the actual data is usually incomplete. Therefore, the study of the classifiers for incomplete data is of great significance [1-4]. The current classification method commonly used decision tree, neural network, KNN, SVM and the Bayesian methods, KNN algorithm in classification accuracy is simple and high in the Chinese automatic text classification has been applied widely. However, this method is more efficient to reduce the efficiency of the training samples with the large amount of computation, and the result is not very good [5].

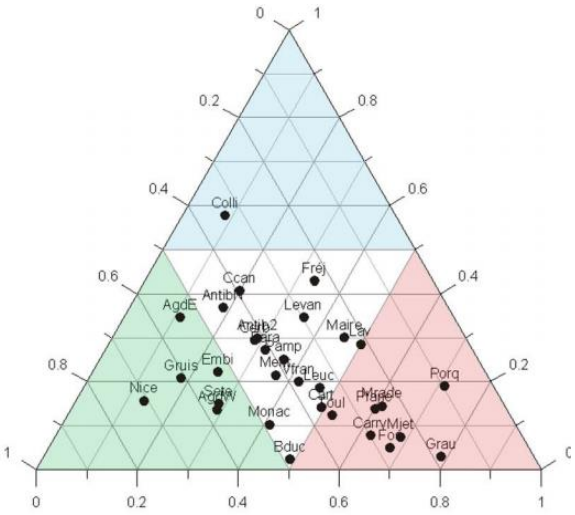


**Figure 1.The data classification benchmark demonstration**

As demonstrated in the figure one, we show the sample benchmark for analysis. After research discovery, if before using the KNN algorithm to the sample attribute reduction, deletes these to the classified result influence small attribute, then we can obtain to treat the classified sample with the KNN algorithm fast the category and can thus obtain better effects [6]. Therefore, the objective function can be defined as follows.

$$\min J_m(U,V) = \sum_{i=1}^{n}\sum_{j=1}^{c} u_{i,j}^{m} d_{i,j}^{2} \quad (1)$$

Where the restriction condition can be defined as formula 2 and 3, respectively.

 

$$\sum_{j=1}^{c} u_{ij} = 1 \tag{2}$$

$$\sum_{j=1}^{n} u_{ij} > 0 \qquad i \le j \le c \tag{3}$$

Random theory when being used in the attribute reduction of decision table may in maintaining the decision table under decision power invariable the premise while deletes not related redundant attribute. We can start our algorithm by considering this: First calculates to treat the classified sample and known category distance or the similarity of between the training samples while found the distance or the similarity with treating the classified sampled data recent K neighbors that judges to treat the classified sampled data according to these neighbor respective categories of again the category. K neighbors who if treats the classified sampled data is a category, then waits the classified sample also to be this category [7-8]. Besides this, we can consider using the kernel to optimize [9]. In the formula 4, we define the distance measurement standard.

$$d_{Distance} = \sqrt{\sum_{i=1}^{n} \left( x_i - y_i \right)^2} \tag{4}$$

Where the $x_i$ and $y_i$ subject to the following vectors.

$$X = \left( x_1, x_2, x_3, \cdots, x_n \right)^2 \tag{5}$$

$$Y = \left( y_1, y_2, y_3, \cdots, y_n \right)^2 \tag{6}$$

In order to avoid the above-mentioned Naive Bayes classifier and the RBC classifier problems, in the next part of the DBCI classifier is given. A is not missing, C is missing, A and C are missing three kinds of statistics, and these frequencies according to the A and C values of the missing variables, including A missing, and the C is not missing. The frequency of each observation of C is proportionally allocated to the respective frequency. In Table 1, we show the general steps of the method.

**Table 1. The general steps of the proposed classification method**

| Procedure | Operation Guidance |
|---|---|
| ***Data Collection*** | By using the common data model to encapsulate the local shared data, the internal structure is hidden, and the uniform public external access interface is provided in the XML format, which shields the different database systems from the different database systems of the data source location. |
| ***Feature Select*** | The classified decision-making of Bayes classifier in feature space using statistical method, when distinguishes the object to converge the specific type. The core technologies determine a ruling rule in the sample training regulations foundation. |
| ***Classification*** | Goal of Bayesian classification is through the machine learning function the heterogeneous data<br>Attribute value of record in storehouse according to data model and record classifies to advance the domain ontology category of definition forms the knowledge node of knowledge library. |

In the classification result test, two-tailed t tests with a significance level of 95% were used to test whether the two classifiers had a significant difference in the accuracy of the classification on each data set. Significantly high accuracy on each data set is shown in

the bold. If there is no significant difference in the classification accuracy of the two classifiers on a data set and they are in bold. The later experiment will verify this.

### 2.2. Data Mining Model Review

Data mining is the process of extracting the hidden and potentially unknown and the potentially useful information and knowledge from a large number of the structured and unstructured data. It requires that the data source should be a large number of real as multimedia, the discovery and extraction of information and knowledge is latent, effective and hidden behind a large amount of data, the user is interested in, understandable and can use the knowledge, data mining is a variety of analytical tools in the mass as the process of discovering the relationships between models and data in data. In the following Figure 2, we show the organization of the data mining technology.

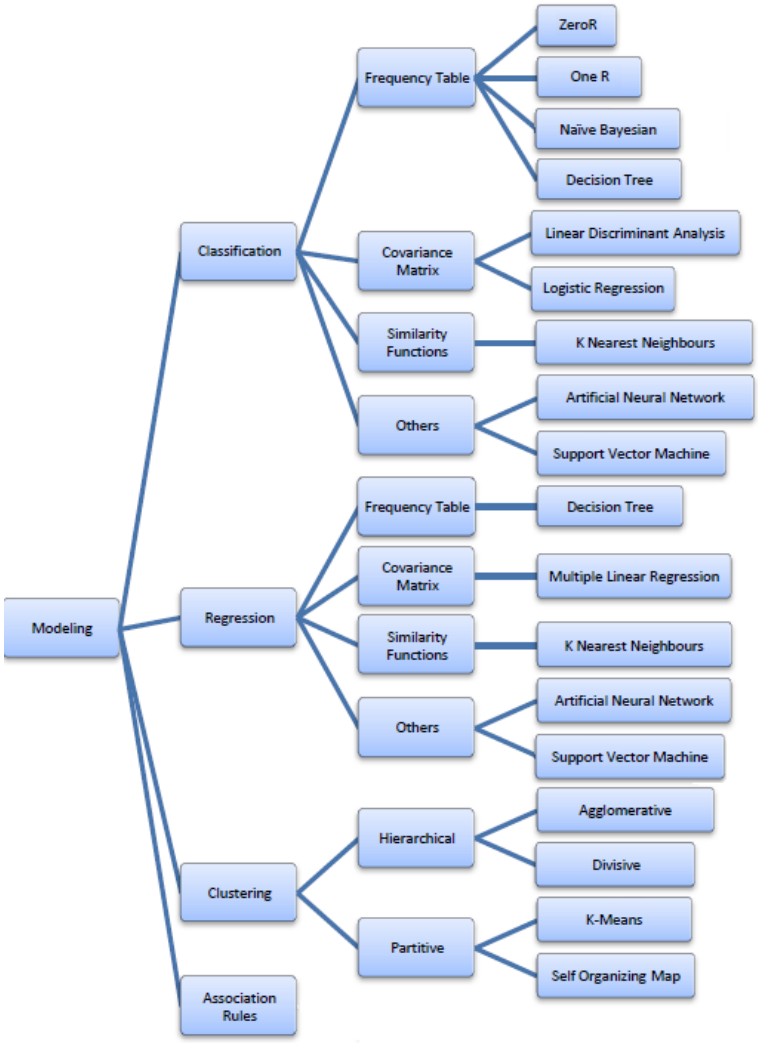


**Figure 2. The organization of the data mining technology**

Among all the method, the basic connection discovery method is the essential one. Compared with the traditional data mining method, the relation discovered that has the unique elements in 5 aspects: (1) The data is the isomerism, from many origins, including: people, events, object, movement, plan and the organization. (2) In the relation discovery, node expressed that the entity relates is among them the relations, in the traditional data mining method, the node expressed that the variable relates expressed the probability of

 

among variables relating. (3) The probability that the relation discovery must determine special graph theory structure-based the data example and is worth between the patterns that pays attention to match. (4) All relations found that the problem appraises the entire community through the data of sampling, but sample quantity generally few. (5) The data value will drain along with time, when therefore the crucial question of relation discovery is to choose makes the decision and set up the corresponding paradigms.

## 3. The Proposed Correlation Analysis Methodology

The gray incidence refers to among the things does not determine the connection, or between the system factors, factors during main behaviors does not determine the connection. The basic task of the analytical grey incidence is the microscopic or macroscopic geometry of behavior based factor sequence is close, to analyze and contribution degree of influence or the factor between determination factors to main behavior, but the gray incidence space carries on the foundation of analytical grey incidence. The Figure 3 shows the corresponding architectures.

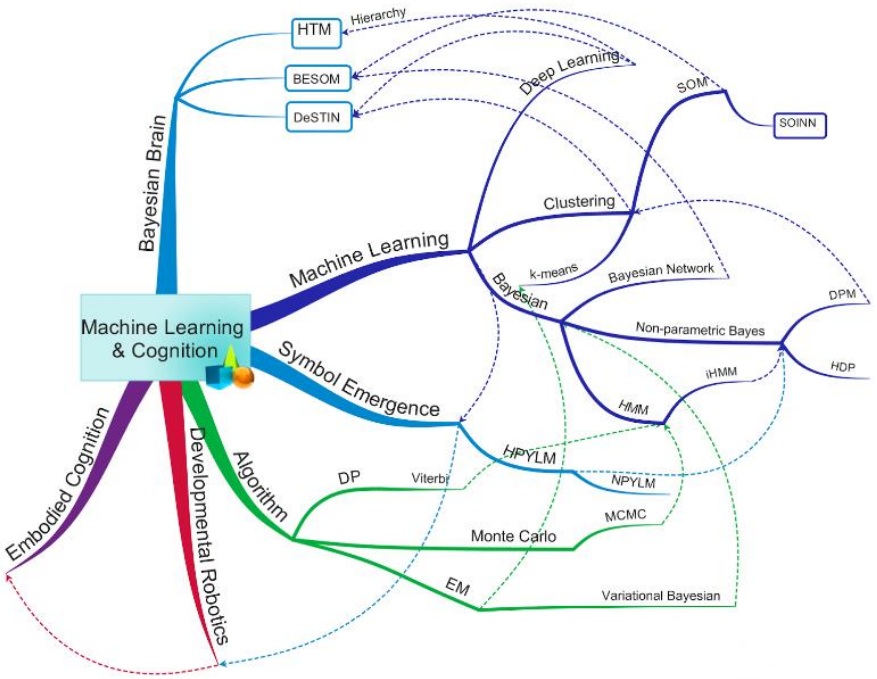


**Figure 3. The corresponding architectures correlation analysis model**

Gray relational analysis, which is the factor analysis of systems is the method to analyze the correlation degree of each factor in the system. The basic idea is to determine the degree of similarity between the reference sequence and the number of comparison sequences to determine whether it is close, reflecting the relationship between the curves. The degree of association among the factors is measured by the degree of correlation. The correlation degree of comparison sequence to a reference sequence is defined as the formula 7.

$$\gamma_i = \frac{1}{n}\sum_{i=}^{n}\xi_i(k) \tag{7}$$

Here, we take the objection projection as the example to analyze that can be then organized as the following aspects. (1) Extensibility and maintainability: Sometimes object-oriented programs need to be prototyped, so that the relationship between

your objects and objects can change frequently, which requires that the mapping methods have extensibility and maintainability. (2) Query processing: sometimes the system requires the data to have the best query performance, such as online real-time processing system; sometimes the system requires no data redundancy, and in a specific format organization to adapt the special query. (3) Performance: read, write and update the performance, the key is to read, write and update the data from the database to read, write and update the number of data required to access the table. According to the different mapping methods, sometimes an object needs to be mapped into a relational table, sometimes to be mapped into multiple relational tables, which leads to the difference between reading and the writing and updating performance, sometimes the same mapping method, it has read the relatively good performance, a new performance may also be poor. (4) Storage redundant: refers to the relationship with the object map data is stored in a table, some mapping method difference redundancy, and some methods may produce number of redundancy.

Before calculating correlation coefficient and the raw data processing because the reference sequence or comparing values given unit and magnitude in general is not the same, so that we have to do the original data dimensionless processing. The gradation correlation analysis can judge the similarity between the two gray systems compare. Regarding the gray system appraisal issue of mapping approach choice, the unit process to be equipped with m treats the appraisal mapping approach, the evaluation standard for all target values of each mapping approach indicated with the vector. The corresponding factor can be then defined as formula 8.

$$\theta_{balance} = \frac{\min\limits_{i}\min\limits_{j}\left|x_{0j}-x_{ij}\right| + \max\limits_{i}\max\limits_{j}\left|x_{0j}-x_{ij}\right|}{\left|x_{0j}-x_{ij}\right| + R\max\limits_{i}\max\limits_{j}\left|x_{0j}-x_{ij}\right|} \tag{8}$$

What the incidence coefficient $\theta_{balance}$ reflection are the evaluation scheme and the connection in the synergy target, to reflect the synthesis connection degrees of all targets, we must carry on the weight sum to obtain the interrelatedness of various plan and optimal alternate according to the development levels of various evaluation standards to some of the various incidence coefficients. There are a lot of data on the correlation coefficient, so the information is too scattered and inconvenient to compare. For this reason, it is necessary to focus the correlation coefficient of each moment as the numerical value and find out the average value to deal with the information. This is the correlation degree, which reflects the comparison sequence and relevancy of the reference sequence defined as formula 9 [10-11].

$$\Delta_{finalized} = \frac{\Delta Min + \rho\Delta Max}{\Delta_i(k) + \rho\Delta Max} \tag{9}$$

## 4. The Proposed Environmental Quality Evaluation Model

### 4.1. Environmental Quality Modelling

In our model, we take the optimization of the domestic waste transportation planning as the example. Reasonable plan of the home scrap logistics transportation not only can effectively reduce the home scrap logistics transportation cost, enhances the home scrap the recycling use factor, then achieves the environmental protection the goal, but can also further promote the resident and support and coordination of the home scrap reclaim and utilize enterprise and the logistics enterprise and ensure the home scrap transportation optimization that causes various work smooth implementation.

 

The research of basic logistics transportation plan issue mainly includes the logistics facilities location and some transportation plan formulates two contents. In the thread of selected location model as most scholars consider the total cost minimum merely that has actually neglected the effect of the logistics transportation on the environment and only slightly mentioned in thread of the transportation plan formulation to the consideration of environmental factor, but did not have the thorough quantification studied. Therefore, as considering that the home scrap logistics shipping center existing home scrap turnover ability and effect of the home scrap transportation time on the environment and proposed a type and quantity method to environmental effect this variable. We firstly define the core impact of the general environment as the formula 10.

$$F_{influence} = \begin{cases} 1, & x \le x_{\max} \\ 1 - \left( \dfrac{x_i - x_{\max}}{\alpha x_{\max}} \right) & x > x_{\max} \end{cases} \tag{10}$$

Expressed when some home scrap students send out the order request, after home scrap transportation to keeping in the general stock center, the shipping center can immediately classification-based treatment this batch of trash, and ships to the recovery plant or the landfill promptly, thus meets the requirements of environmental protection and the basic environmental impact factor evaluation can be summarized as the follows.

$$F_{total\_impact} = \frac{\sum_i F_i(x_i) x_i}{\sum_i x_i} \tag{11}$$

The goal of the waste transport planning problem is to minimize the total cost of the logistics system in the limited storage capacity of the transport centers and minimize the environmental impact associated with the transport capacity and the transport time of the existing network, a multi-objective optimization model for the planning of core logistics center with three goals is established. The objective function is defined as below.

$$\min Y(\delta_j, \delta_{ijk}) = \sum_j C_{fixj} \delta_j + \sum_J V_j \left( \sum_i \sum_k D_{ki} \delta_{ijk} \right)^{\theta} \tag{12}$$

Each objective function is a core 0-l integer programming problem, and each function usually has conflicting properties. In the solution, the global optimal solution cannot be generalized, and only the objective function cannot be obtained that can be drawn non-inferior solution set, the collection of elements is not dominant. To solve optimization problem, we use the prior discussed methodology to achieve the goal [12-15].

### 4.2. Environmental Quality Evaluation and Analysis Discussion

In this sub-section, we take the air and water quality evaluation as the sample to apply the proposed method for verification. The appraisal significance of air quality is great, for the past dozen years, the evaluation study of atmospheric environment quality has made the noticeable progress has been used in the appraisal of atmospheric environment quality besides the exponential method and PCA law and level decision-making law, the fuzzy set theory and the gray system analysis, along with the development of establishment and computation technology of some new disciplines, domestic and foreign also proposed the new method of many atmospheric environment quality appraisal. As part of our model, we apply the BP neural network to verify the issue. The BP network can study and store massive inputs-output mode mapping relationship, but did not need to reveal to describe the mathematical model of this mapping relationship in advance.

For the air quality assessment, the BP network is composed of a hidden layer. According to GB3095-2000, the input vector of the training sample contains three components, three components correspond to the concentration of three pollutants,

the output vector contains a component and the component represents the true level. The number of nodes in the hidden layer has some influence on the performance of the neural network. The number of nodes is too small to meet the required accuracy of error. Too many nodes often increase the learning time, and the trained network model is prone to over-fitting and the loss of generalization ability. At present, the determination of the number of nodes in the hidden layer is generally based on the following empirical formula to determine a range as follows.

$$l = \sqrt{n+m} + a \tag{13}$$

Regarding batch run way, the iterative time that we establish is 10000 times, the learning rate when decision causes the network can stabilize studies the maximum value, this can pick up the learning speed generally, certainly regarding the VLBP algorithm, because of the learning rate can auto-adapted, therefore chose value not to relate. We discovered that regarding 5 algorithms, network model of establishes after 3 minutes of 10,000 iterative learning can be 100% rates of accuracy regarding 30 test samples in the best situation, certainly, because initialization each time is different, therefore trains after each time, performance of network has difference, this is mainly because fell into the different minimum points, we tested repeatedly to find overall situation most dot.

Regarding the smooth learning mode, we establish the training sample circulation to study 200 times, the time that such study spends iterates the time that 10000 times spend almost to be the same with the batch run way, after this learning mode training is completed will not output the mean error, after we can the programming of calculation training complete, tests the sample the mean error. Then, we analyze the water quality evaluation. The development of the fuzzy mathematics theory has opened the way for the application of fuzzy theory in the classification of water quality evaluation.

As a result of the comprehensive utilization of water quality, the comprehensive utilization of water quality is becoming more and more important, and it is difficult to use simple mathematical model to reflect the various pollutants and environment. In the evaluation, the pollution degree and the water quality classification boundary are some fuzzy concepts and the phenomena which exist objectively. Therefore, it is better to use the fuzzy set theory to classify and evaluate the water quality than the comprehensive index method to reflect objective condition of environmental quality. Regarding the gray cluster of hydrology environment quality, what the cluster object refers to is the contaminated area (*e.g.*, sea and the river mouth district). (Upstream, middle reaches and downstream how the contamination control cross section flows) or pollution time interval (*e.g.*, years). The cluster target is the contamination index that to refer to giving, correspondingly cluster target from the number is the density of contamination index. The cluster ash spreads usually refer to the water pollution degree graduation. According to the basic concept of grey active cluster, this article proposed is applicable to the multi-objects and the multi-objectives, grey number and gray cluster method of hydrology environment quality is as later section shows.

### 4.3. Results Analysis

In this sub-section, we conduct the numerical analysis. The sampling process is no exception and the experiment shows that the sampling error is often the largest and most important error. Check the sampling of all aspects of the problems that occur in the field sampling at the time of the sampling. Sampling quality control samples to determine the source of the error and sampling at the same sampling point, while collecting two copies of parallel samples, delivered to the laboratory by the password analysis, this method is

 

simple, the drawback is that only sampling and analysis of the whole process of precision. Analysis of quality control includes many steps as each step has some control standards, such as blank, detection limit, sensitivity curve boundaries and determination of standard material allowable error, and among the many indoor measured values between allowable errors. At first, the multiple evaluation criteria by method of "business", according to the experience needs and possibilities for error analysis determined to set the absolute error and the relative error allowed range. But because it has no strict definition and calculation method, and therefore the judgment made this judgment with confidence how much the probability cannot explain whether under control. Therefore, in recent years, most of the methods the mathematical statistics to formulate the various control standards. The most widely used method is using the collaborative experiment making analysis of allowable deviation. Under this basis, we design the data extraction system as the Figure 4. And the numerical simulation results are shown in the Figure 5 and 6, respectively.

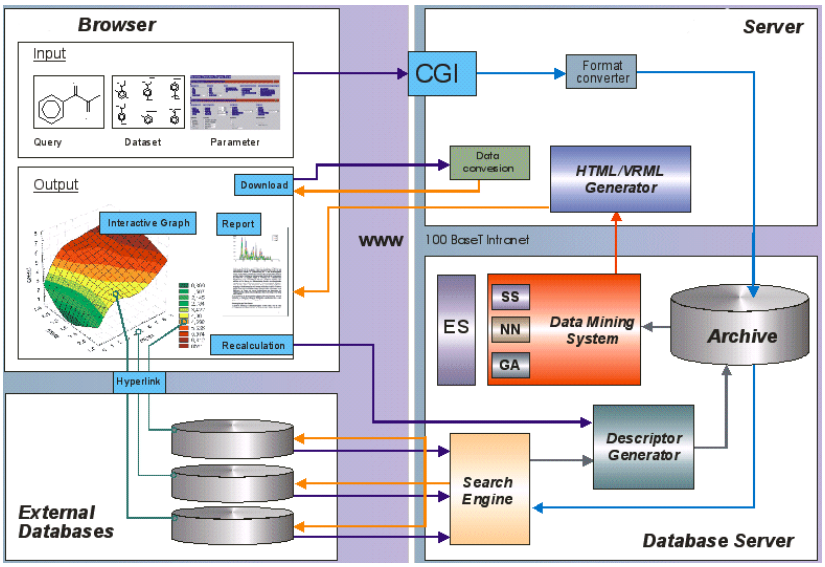


**Figure 4. The systematic architecture on the proposed method**

| | Environmental parameters | | | | | | | Foraminiferal parameters | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Water depth | %OM | %<63µm | %63-125µm | %125-250µm | %250-500µm | %>500µm | %tolerant sp. | %sensitive sp. | %epiphytic sp. | %perforate sp. | %porcelaneous sp. | %agglutinated sp. | Absolute densities | Specific richness | Shannon index | Equitability index | ES50 | %Etstd |
| **Environmental parameters** | | | | | | | | | | | | | | | | | | | |
| Water depth | | 0.07 | 0.42 | 0.08 | 0.15 | 0.16 | 0.01 | 0.45 | 0.22 | 0.00 | 0.27 | 0.80 | 0.22 | 0.72 | 0.00 | 0.00 | 0.00 | 0.00 | 0.21 |
| %OM | 0.34 | | 0.02 | 0.02 | 0.00 | 0.83 | 0.18 | 0.03 | 0.59 | 0.07 | 0.01 | 0.11 | 0.03 | 0.39 | 0.01 | 0.00 | 0.10 | 0.01 | 0.05 |
| %<63µm | -0.16 | **0.42** | | 0.87 | 0.00 | 0.00 | 0.04 | 0.01 | 0.01 | 0.07 | 0.02 | 0.00 | 0.66 | 0.06 | 0.39 | 0.80 | 0.47 | 0.51 | 0.17 |
| %63-125µm | -0.33 | **-0.42** | -0.03 | | 0.03 | 0.00 | 0.00 | 0.75 | 0.02 | 0.01 | 0.08 | 0.85 | 0.01 | 0.36 | 0.61 | 0.07 | 0.15 | 0.22 | 0.60 |
| %125-250µm | -0.27 | **-0.56** | **-0.60** | **0.41** | | 0.61 | 0.03 | 0.07 | 0.84 | 0.11 | 0.00 | 0.00 | 0.09 | 0.40 | 0.28 | 0.08 | 0.18 | 0.17 | 0.28 |
| %250-500µm | 0.27 | -0.04 | **-0.59** | **-0.62** | 0.10 | | 0.01 | 0.06 | 0.00 | 0.03 | 0.46 | 0.00 | 0.10 | 0.11 | 0.98 | 0.22 | 0.16 | 0.22 | 0.21 |
| %>500µm | **0.50** | 0.25 | **-0.39** | **-0.63** | **-0.39** | **0.46** | | 0.39 | 0.00 | 0.00 | 0.10 | 0.54 | 0.00 | 0.16 | 0.93 | 0.11 | 0.03 | 0.04 | 0.67 |
| **Foraminiferal parameters** | | | | | | | | | | | | | | | | | | | |
| %tolerant sp. | 0.15 | **0.40** | **0.48** | 0.06 | -0.34 | -0.35 | -0.17 | | 0.01 | 0.55 | 0.01 | 0.00 | 0.37 | 0.40 | 0.10 | 0.14 | 0.68 | 0.55 | 0.00 |
| %sensitive sp. | 0.23 | 0.10 | **-0.49** | **-0.44** | -0.04 | **0.59** | **0.60** | **-0.48** | | 0.00 | 0.33 | 0.01 | 0.00 | 0.78 | 0.92 | 0.44 | 0.39 | 0.28 | 0.03 |
| %epiphytic sp. | **0.53** | 0.34 | -0.34 | **-0.50** | -0.30 | **0.40** | **0.80** | -0.12 | **0.65** | | 0.01 | 0.95 | 0.00 | 0.24 | 0.48 | 0.01 | 0.00 | 0.01 | 0.92 |
| %perforate sp. | 0.21 | **0.50** | **0.42** | -0.33 | **-0.63** | -0.14 | 0.32 | **0.48** | 0.19 | **0.46** | | 0.00 | 0.00 | 0.06 | 0.44 | 0.19 | 0.19 | 0.25 | 0.03 |
| %porcelaneous sp. | -0.05 | -0.30 | **-0.71** | -0.04 | **0.57** | **0.55** | 0.12 | **-0.52** | **0.50** | -0.01 | **-0.65** | | 0.88 | 0.06 | 0.51 | 0.54 | 0.79 | 0.89 | 0.04 |
| %agglutinated sp. | -0.24 | **-0.40** | 0.09 | **0.47** | 0.32 | -0.31 | **-0.52** | -0.17 | **-0.69** | **-0.60** | **-0.74** | -0.03 | | 0.43 | 0.67 | 0.25 | 0.13 | 0.16 | 0.30 |
| Absolute densities | -0.07 | 0.17 | 0.36 | 0.18 | -0.16 | -0.30 | -0.27 | 0.16 | 0.05 | -0.22 | 0.35 | -0.35 | -0.15 | | 0.04 | 0.95 | 0.01 | 0.63 | 0.61 |
| Specific richness | **0.66** | **0.46** | 0.17 | -0.10 | -0.21 | 0.01 | 0.02 | 0.31 | 0.02 | 0.14 | 0.15 | -0.13 | -0.08 | **0.38** | | 0.00 | 0.16 | 0.00 | 0.07 |
| Shannon index | **0.74** | **0.52** | 0.05 | -0.34 | -0.33 | 0.23 | 0.30 | 0.28 | 0.15 | **0.45** | 0.25 | -0.12 | -0.22 | -0.01 | **0.82** | | 0.00 | 0.00 | 0.09 |
| Equitability index | **0.52** | 0.31 | -0.14 | -0.28 | -0.26 | 0.27 | **0.42** | 0.08 | 0.17 | **0.57** | 0.25 | -0.05 | -0.29 | **-0.48** | 0.27 | **0.69** | | 0.00 | 0.52 |
| ES50 | **0.79** | **0.47** | -0.13 | -0.24 | -0.27 | 0.24 | **0.39** | 0.12 | 0.21 | **0.51** | 0.23 | -0.03 | -0.28 | -0.10 | **0.81** | **0.97** | **0.82** | | 0.33 |
| %Etstd | 0.24 | 0.36 | 0.26 | 0.10 | -0.21 | -0.24 | -0.08 | **0.97** | **-0.40** | -0.02 | **0.41** | **-0.38** | -0.20 | 0.10 | 0.34 | 0.32 | 0.12 | 0.193 | |

**Figure 5. The statistical simulation result on the air quality**

| | Station | OPD (mm) | ALD5/ ALD6 (cm) | Depth (m) | %OM | %<63μm | %63-125μm | %125-250μm | %250-500μm | %>500μm |
|---|---|---|---|---|---|---|---|---|---|---|
| **Group A** | | | | | | | | | | |
| | Grau | 7 | 1.0 | 15 | 3.28 | **67.92** | 25.86 | 6.21 | 0.00 | 0.00 |
| | Toul | - | 1.1 | 43 | 7.52 | **50.90** | 11.59 | 4.84 | 2.82 | 29.85 |
| | Cart | 6 | 1.3 | 10 | 5.91 | **80.80** | 13.07 | 6.14 | 0.00 | 0.00 |
| | Mjet | 14 | 1.4 | 41 | 5.41 | **51.91** | 19.80 | 19.78 | 7.61 | 0.90 |
| | Nice | 12 | 1.4 | 30 | 2.21 | **46.12** | **28.47** | 17.79 | 6.80 | 0.81 |
| | Bduc | - | 1.6 | 14 | 1.68 | **87.52** | 10.75 | 1.73 | 0.00 | 0.00 |
| **Group B** | | | | | | | | | | |
| | Colli | - | 1.4 | 23 | 1.37 | 2.89 | 13.86 | **39.90** | **33.74** | 9.62 |
| | AgdE | - | 1.8 | 21 | 1.57 | 8.45 | 27.56 | **56.99** | 6.99 | 0.00 |
| | Maire | - | 2.0 | 40 | 3.31 | 4.82 | 4.26 | 11.22 | 23.28 | **56.42** |
| | Vfran | - | 2.0 | 42 | 3.99 | 14.66 | 12.25 | 18.11 | **22.75** | **32.22** |
| | Leuc | - | 2.2 | 22 | 1.72 | 19.73 | **53.10** | 24.61 | 2.41 | 0.14 |
| | Carry | - | 2.3 | 48 | 3.54 | **26.28** | 14.25 | 17.57 | 17.72 | **24.18** |
| | Ment | - | 2.3 | 51 | 1.73 | 28.26 | **37.50** | **32.75** | 1.49 | 0.00 |
| | Pamp | - | 2.7 | 42 | 2.78 | 19.10 | 6.06 | 11.39 | **23.95** | **39.49** |

**Figure 6. The statistical simulation result on the water quality**

## 5. Conclusions

In this paper, we conduct research on environmental quality evaluation model based on data mining and correlation analysis. Because the environment system is an open system, the change of environmental quality is various results of aggregation of variable function. Along with the application of multi-statistical analysis method, the big data analysis law by has been applied in the environmental quality evaluation. Reciprocities of this method among from many targets starts that changes into a few not related overall targets many targets and the merit lies in had considered the relevance among various targets that can maximum limit retain original information, carries on best comprehensive dimensionality reduction processing to the high dimensional data, and determined that objectively the weight number of each target which has avoided the subjective randomness. Based on this feature, we propose the data mining and correlation analysis based model. The proposed model performs well according to experimental analysis. In the future, more simulation will be down to test the systematic robustness.

## Acknowledgments

The work of this paper is supported by the National Natural Science Foundation of China (grant number 31460090).

## References

[1] D. Liu and Z. Zou, “Water quality evaluation based on improved fuzzy matter-element method”, Journal of Environmental Sciences, vol. 7, (**2012**), pp. 1210-1216.
[2] C. Andrade, M. L. Lima, F. Fornara and Marino Bonaiuto, “Users' views of hospital environmental quality: Validation of the perceived hospital environment quality indicators (PHEQIs)”, Journal of environmental psychology, vol. 32, no. 2, (**2012**), pp. 97-111.
[3] A. Yalcuk and S. Postalcioglu, “Evaluation of pool water quality of trout farms by fuzzy logic: monitoring of pool water quality for trout farms”, International Journal of Environmental Science and Technology, vol. 12, no. 5, (**2015**), pp. 1503-1514.
[4] L. A. Manfré, A. M. Da Silva, R. C. Urban and J. Rodgers, “Environmental fragility evaluation and guidelines for environmental zoning: a study case on Ibiuna (the Southeastern Brazilian region)”, Environmental earth sciences, vol. 69, no. 3, (**2013**), pp. 947-957.
[5] J. M. Ronan and B. McHugh, “A sensitive liquid chromatography/tandem mass spectrometry method for the determination of natural and synthetic steroid estrogens in seawater and marine biota, with a focus on proposed Water Framework Directive Environmental Quality Standards”, Rapid Communications in Mass Spectrometry, vol. 27, no. 7, (**2013**), pp. 738-746.
[6] A. Sarkar, J. Bhagat and S. Sarker, “Evaluation of impairment of DNA in marine gastropod, Morula granulata as a biomarker of marine pollution”, Ecotoxicology and environmental safety, vol. 106, (**2014**), pp. 253-261.
[7] G. Merrington, Y. -J. An, E. P. M. Grist, S. -W. Jeong, C. Rattikansukha, S. Roe and U. Schneider, “Water quality guidelines for chemicals: learning lessons to deliver meaningful environmental metrics”, Environmental Science and Pollution Research, vol. 21, no. 1, (**2014**), pp. 6-16.

 

[8] L. Paoli, A. Corsini, V. Bigagli, J. Vannini, C. Bruscoli and S. Loppi, "Long-term biological monitoring of environmental quality around a solid waste landfill assessed with lichens", Environmental Pollution, vol. 161, (**2012**), pp. 70-75.
[9] H. Wang and J. Wang, "An effective image representation method using kernel classification", In 2014 IEEE 26th International Conference on Tools with Artificial Intelligence, IEEE, (**2014**), pp. 853-858.
[10] P. Sengupta, T. Jovanovic-Talisman and J. Lippincott-Schwartz, "Quantifying spatial organization in point-localization superresolution images using pair correlation analysis", Nature protocols, vol. 8, no. 2, (**2013**), pp. 345-354.
[11] Z. Yuan, J. Li, J. Li, X. Gao and S. Xu, "SNPs identification and its correlation analysis with milk somatic cell score in bovine MBL1 gene", Molecular biology reports, vol. 40, no. 1, (**2013**), pp. 7-12.
[12] M. U. Gutmann, V. Laparra, A. Hyvärinen and J. Malo, "Spatio-chromatic adaptation via higher-order canonical correlation analysis of natural images", PloS one, vol. 9, no. 2, (**2014**), pp. e86481.
[13] X. Wu, X. Zhu, G. -Q. Wu and W. Ding, "Data mining with big data", IEEE transactions on knowledge and data engineering, vol. 26, no. 1, (**2014**), pp. 97-107.
[14] M. A. Vieira, A. R. Formaggio, C. D. Rennó, C. Atzberger, D. A. Aguiar and M. P. Mello, "Object based image analysis and data mining applied to a remotely sensed Landsat time-series to map sugarcane over large areas", Remote Sensing of Environment, vol. 123, (**2012**), pp. 553-562.
[15] R. Vijayakrishnan, S. R. Steinhubl, K. Ng, J. Sun, R. J. Byrd, Z. Daar, B. A. Williams, S. Ebadollahi and W. F. Stewart, "Prevalence of heart failure signs and symptoms in a large primary care population identified through the use of text and data mining of the electronic health record", Journal of cardiac failure, vol. 20, no. 7, (**2014**), pp. 459-464.

 