SSD Model Selection Method Based on Machine Learning Algorithm

Bing Yang1, Hong Yang2, Chengmei Zhang1, Yajie Wang1 and Miao Hao1,2,*
1 Guizhou Academy of Testing and Analysis, Guiyang, Guizhou, China
2 Guizhou Academy of Sciences Big Data Co.LTD, Guizhou, China
* Corresponding author: haomiao@gzbdi.com

Abstract. The chemical pollutants produced by human activities often lead to the risk of ecological species diversity destruction by affecting the physiological activities of organisms in nature. Ecological risk assessment (ERA) can be used to evaluate the degree of adverse effects of external factors on the ecological environment, provide the basis for taking effective ecological protection measures and formulating reasonable environmental policies, and also an important means to understand the possible adverse effects of ecological health and pollutants on the ecological environment. The species sensitivity distribution (SSD) method is a widely used evaluation method, and its core step is to select the appropriate species toxicity data for curve fitting. In this paper, the basic concept of ecological risk assessment is briefly introduced, and the basic principle and implementation steps of species sensitivity distribution method are described in detail. Aiming at the problem of SSD model selection in water environment, machine learning algorithm is introduced, and corresponding neural network is constructed. Taking the root mean square (RMSE) and sum of squared error (SSE) were as the index, the optimal SSD model suitable for water environment is determined, The SSD model selection method based on machine learning algorithm is obtained. At last, the application of machine learning algorithm in SSD model selection is prospected, and the shortcomings of the existing research and the future research direction are pointed out.

1. Introduction
Ecological risk refers to the risk that the ecosystem and its components bear. It usually refers to the possible adverse effects of pollutants and other stress factors on the ecosystem and its components, including the damage to the structure and function of the ecosystem. The receptor object of ecological risk is a complex system, including species of multiple trophic levels in the ecosystem. For the aquatic ecosystem, there are algae, aquatic vascular plants, zooplankton, benthos, fish, bacteria and so on [1-2].

Ecological risk assessment is a process to evaluate the adverse ecological effects of one or more external factors on aquatic ecosystem. Ecological risk assessment is to use the comprehensive knowledge of environmental science, ecology, geography, biology and other disciplines, using mathematical, probability theory and other quantitative analysis techniques to predict, analyze and evaluate the possible damage to aquatic ecosystem and its components caused by uncertain events. To sum up, ecological risk assessment refers to the assessment of the possibility and harmfulness of adverse ecological consequences after the aquatic ecosystem is affected by one or more stress factors, so as to put forward countermeasures and measures to prevent or reduce adverse effects and ensure the safety and health of the ecosystem [3].
The species sensitivity distribution (SSD) method is based on the difference of sensitivity of different species to the same pollutant. It is based on the acute or chronic toxicity data of several representative sensitive species, and uses the toxicity data value or conversion value as abscissa and the cumulative probability of toxicity data as ordinate to construct the statistical distribution model. Then the distribution model is built to calculate the proportion of environmental concentration of species of the ecosystem [4-5]. This method is suitable for the situation with more data and less uncertainty, and can quantitatively express the evaluation results, taking into account the species diversity and the integrity of the ecosystem, so it is widely used in ecological risk assessment. SSD method is the most widely used method for ecological risk assessment, and its advantage is that it can predict the potential ecological effects of pollutants in the ecosystem from the results of single species laboratory toxicity experiments. It models the sensitivity of different species and estimates the potential impact proportion of species that will be harmed when exposed to pollutants. It has good sensitivity and accuracy. However, the traditional SSD method often needs to screen out the best fitting function through a large number of tests in practical application, which leads to the low efficiency of the final ecological risk assessment. Therefore, aiming at this problem, this paper introduces machine learning method to optimize the selection of SSD model and improve the efficiency of ecological risk assessment.

2. Overview of SSD method
The toxicological effects of different trophic organisms on risk sources in ecosystems are different. In toxicology, different species have different dose-response relationships to the same pollutant. That is to say, in a complex ecosystem, the sensitivity of different species to a certain harmful factor obeys a certain cumulative probability distribution. In order to protect the whole ecosystem, a group of representative sensitive species is selected to represent the sensitivity of the whole ecosystem. It is assumed that the sensitivity characteristics of this group of species also obey a certain cumulative probability distribution, and the available species toxicological data is considered to be a sample from this distribution. Therefore, the concentration values of acute or chronic toxicity data of different organisms are plotted according to the quantiles arranged in size, and a distribution function is selected to fit the parameters of these points, and the SSD curve is obtained. The concentration level of pollutants to protect most species in an ecosystem is determined by SSD curve, and 5% HC5 (Hazardous Concentration 5%) is generally used. HC5 refers to the concentration of 5% species in the ecosystem, which can be compared with the concentration of known pollutants to determine whether there is ecological risk. At the same time, the potential impact ratio PAF can be calculated by SSD curve under the condition of known pollutant concentration to characterize the ecological risk. According to the principle of species sensitivity distribution method, SSD curve method is the basic step of ecological risk assessment

There are five steps: (1) collecting and screening data (2) The fitting model was determined (3) SSD curve was constructed (4) PAF and HCs were calculated (5) Ecological risk assessment. Among them, data collection and screening, fitting model determination are important steps to affect the accuracy of ecological risk assessment results of SSD method.

When constructing SSD curve, the more species with different trophic levels are selected, the more representative the whole ecosystem is. In fact, the results of ecological risk assessment by SSD curve method are more accurate. On the contrary, the less species there are, The greater the uncertainty of the evaluation results. Therefore, there is minimum toxicity data requirement (mtdr) in SSD curve construction to ensure the accuracy of evaluation results. The application of SSD in ecological risk assessment is mainly based on a set of laboratory ecotoxicity data. The maximum concentration of acceptable contaminants, so the effectiveness of SSD depends on the quality and relevance of the data used. When there is no necessary toxicity data, some mature prediction models can be used to obtain toxicity data.

According to the quantity of pollutants, the fitting models used to construct SSD curve can be divided into two categories. One is for single pollutant, the other one is the fitting model of pollutants, and the other is the combined model for multiple pollutants. There are two main methods to construct SSD curve
model of single pollutant: parametric model based on deterministic probability and nonparametric model based on sampling distribution. Among the methods based on deterministic probability model, the main parameter models are log normal, log triangular, log logistic, log Gumbel, Weibull and Burr III. The nonparametric model methods based on sampling distribution mainly include bootstrap method and Monte Carlo method.

Among these models, log triangular model fits the toxicity data well, and the US water quality benchmark is also derived based on this model. Burr III model is more flexible and can approximate many common distributions under certain conditions. Many studies have found that log logistic model fits well with the toxicity data of many pollutants, including heavy metals, pesticides and antibiotics. In addition to the model itself, the amount of data is also an important factor affecting the fitting effect. The less the amount of data, the worse the stability of the fitting effect. When the amount of data is small, multiple models should be selected for fitting, and then the fitting model with the largest correlation coefficient should be selected to reduce the error of the model itself.

3. Introduction of machine learning method

In order to achieve the optimal selection of SSD model, this paper introduces the classical machine learning method BP neural network algorithm for SSD model selection. BP neural network is a multilayer feedforward neural network based on error back propagation algorithm. It consists of three layers, including input layer, hidden layer and output layer.

BP neural network algorithm is a kind of supervised learning mechanism for forward propagation of working signal and reverse propagation of error signal. In the process of forward propagation of the working signal, the information of the input layer is transmitted to the output layer through layer by layer operation of the hidden layer. When the forward propagation of the working information is completed, the difference between the output value and the target value is calculated. If there is an error in the comparison, the error signal needs to be recorded and input backward, and the weights of the neurons need to be modified. This process is repeated many times until the output reaches the desired goal.

In the above iterative process, the error value, that is, the iterative parameter , is very important, and represents the size of the node error. The larger the s value is, the larger the connection weight is. s is defined as

\[ F'(\text{NET}_{pk}) = F(\text{NET}_{pk})[1 - F(\text{NET}_{pk})] \]  

(1)

For the hidden layer, s is defined as

\[ \varepsilon_{pk} = \varepsilon_{pk} \sum_{kj} \omega_{kj} \]  

(2)

Where \( \omega_{kj} \) is the connection weight of output layer node k and hidden layer node j. In the opposite propagation of error s, the weight can be corrected by gradient descent method

\[ \Delta \omega_{kj} = \eta \varepsilon_{kj} \]  

(3)

In order to reduce the training time and increase the stability of the model, the corresponding adjustments are as follows:

\[ \Delta \omega_{kj}(n+1) = \eta \varepsilon_{kj}(n) + \alpha[\omega_{kj}(n) - \omega_{kj}(n-1)] \]  

(5)

The final weight is calculated as follows:

\[ \omega_{kj}(n+1) = \omega_{kj}(n) + \Delta \omega_{kj}(n+1) \]  

(6)

The adjustment of connection weight between hidden layer and input layer is a dynamic process. For all samples, when the error reaches the target value or smaller than the target value, the iteration is completed.
4. Experimental test

4.1. Data selection
In this paper, hepatotoxin and neurotoxin were selected as two kinds of algal toxins. Nodurarins and Cylindrospermopsin, Antin-a and Saxitoxins were selected as four typical algal toxins. Sulfonamides, tetracyclines and macrocyclic lipids were selected as three kinds of antibiotics, Sulfamethoxazole (SMX) and Sulfamethazine (SMZ), Tetracycline (TC) and Oxytetracycline (OTC), Erythromycin (ETM) and Roxithromycin (RTM) were used as six typical antibiotics. Because the acute toxicity of microcystins to species in the aquatic environment of China is obvious, the acute toxicological data of aquatic organisms in China are selected for the study; Antibiotics in the water environment pollution concentration is low, the acute toxicity of species is not significant, so the chronic toxicological data of antibiotics were selected for research..

In order to achieve the reliability and efficiency of ecological risk assessment, the selection requirements of biological species for constructing aquatic ecological risk assessment model of microcystins and Antibiotics are as follows: select the typical species in the water environment, which can represent the different trophic levels, trophic types and life forms of the ecosystem, and strive to reflect the whole picture of the effluent ecosystem with the minimum amount of data.

To provide necessary concentration monitoring data for aquatic ecological risk assessment, the pollution level of microcystins and antibiotics in China was investigated by consulting relevant literature. The data screening principles are as follows: (1) the extraction and monitoring methods of pollutants meet the requirements of corresponding standard methods, and the data in the literature with the same detection methods are selected for the same pollutant. (2) the label of pollutant monitoring sites is clear, and the arithmetic mean value of the monitoring data of the same pollutant in the same water area in five years is taken, as shown in Table 1 and Table 2

| Types          | Name | Concentration range (ng·L⁻¹) | Mean concentration (ng·L⁻¹) |
|----------------|------|-------------------------------|----------------------------|
| Sulfonamides   | SMX  | 55.20-765.00                  | 410.48                     |
|                | SMZ  | 89.10-623.00                  | 367.17                     |
| Tetracyclines  | TC   | 114.00                        | 114.00                     |
|                | OTC  | 84.50-219.80                  | 152.15                     |

| Types          | Name | Concentration range (ng·L⁻¹) | Mean concentration (ng·L⁻¹) |
|----------------|------|-------------------------------|----------------------------|
| Sulfonamides   | SMX  | 22.60-40.60                   | 33.60                      |
|                | SMZ  | 218.00                        | 218.00                     |
| Macrolides     | ETM  | 121.00-890.50                 | 379.17                     |
|                | RTM  | 12.00-348.50                  | 180.25                     |

4.2 Experimental result
On the basis of toxicological data, daphnia pulex was used as an example. Five common fitting functions (Burr III, Log normal, Log logistic, Weibull and Gamma) were used to obtain the corresponding SSD curves. The goodness of fit evaluation can be used to test whether the distribution of a kind of data in the population is consistent with a certain theoretical distribution. For parameter model, the main evaluation indexes of goodness of fit are RMSE and SSE. The more the two indexes tend to zero, the better the goodness of fit of the fitting function is. Therefore, this paper takes these two indexes as the core, and uses BP neural network to adjust the model parameters adaptively, The best model was judged by goodness of fit.
The cumulative probability may also affect the goodness of fit of the fitting function. Therefore, set $p \leq 30\%$ as low cumulative probability, $30\% < p \leq 80\%$ as medium cumulative probability, and $P > 80\%$ as high cumulative probability. Compare and analyze the goodness of fit of the five fitting functions under different cumulative probability conditions. The results are shown in Table 3. It can be seen that there are some differences in the goodness of fit of the fitting function under different cumulative probability conditions. Log logistic model is preferred under the condition of low and medium cumulative probability ($0 \leq p \leq 80\%$), and burr III model is preferred under the condition of high cumulative probability ($p > 80\%$).

| Evaluation index | Function   | $p \leq 20\%$ | Sorting | $20\% < p \leq 80\%$ | Sorting | $p > 80\%$ | Sorting |
|------------------|------------|---------------|---------|------------------------|---------|------------|---------|
| SSE              | Log-logistic | $2.42 \times 10^{-4}$ | 1       | 0.017                  | 1       | 0.152      | 2       |
|                  | Gamma      | $2.52 \times 10^{-4}$ | 2       | 0.023                  | 3       | 0.201      | 4       |
|                  | Log-normal | $2.83 \times 10^{-4}$ | 3       | 0.022                  | 2       | 0.179      | 3       |
|                  | Weibull    | $1.45 \times 10^{-4}$ | 4       | 0.027                  | 4       | 0.253      | 5       |
|                  | Burr III   | $5.37 \times 10^{-3}$ | 5       | 0.174                  | 5       | 0.011      | 1       |
| RMSE             | Log-logistic | $4.01 \times 10^{-3}$ | 1       | 0.031                  | 1       | 0.099      | 2       |
|                  | Gamma      | $4.09 \times 10^{-3}$ | 2       | 0.033                  | 3       | 0.124      | 4       |
|                  | Log-normal | $4.45 \times 10^{-3}$ | 3       | 0.032                  | 2       | 0.112      | 3       |
|                  | Weibull    | $8.82 \times 10^{-3}$ | 4       | 0.039                  | 4       | 0.137      | 5       |
|                  | Burr III   | $1.75 \times 10^{-2}$ | 5       | 0.112                  | 5       | 0.029      | 1       |

### 5. Conclusion and expectation

Aiming at the problem of SSD model selection in water environment, this paper introduces machine learning algorithm and constructs the corresponding neural network. Taking root mean square (RMSE) and sum of square error (SSE) as indexes, the optimal SSD model suitable for water environment is determined, and the SSD model selection method based on machine learning algorithm is obtained. Finally, the application of machine learning algorithm in SSD model selection is prospected.

To evaluate the combined ecological risk of multiple pollutants, we need to consider the combined effects of multiple pollutants. However, the variables in this study are relatively single, and there is still a lack of research on the combined risk of microcystins, antibiotics and other pollutants in the water environment. The risk assessment of single pollutant may underestimate the degree of damage to aquatic ecosystem, which makes the selection of SSD model error. Future research can focus on the complex ecological risk caused by the interaction of multiple pollutants in the water environment, optimize the existing machine learning algorithm, and build a more reliable model to improve the realistic credibility of the ecological risk assessment results.

**Acknowledgments**

This work was supported by Guizhou Provincial Science and Technology Projects (qiankehezhicheng [2018]2775).

**References**

[1] Weisheng Guan, Jianjun Yang and Ping Lu, "Study on water quality criteria of Endocrine disrupting chemicals in Weihe River based on species sensitivity distributions theory," 2011 International Symposium on Water Resource and Environmental Protection, Xi’an, China, 2011, pp. 1147-1149.

[2] A. Cord, R. R. Colditz, M. Schmidt and S. Dech, "Species distribution and forest type mapping in Mexico," 2009 IEEE International Geoscience and Remote Sensing
Symposium, Cape Town, South Africa, 2009, pp. V-260-V-263.

[3] B. Chen and J. Zhao, "Species Sensitive Distribution for Aquatic Biota Exposed to DDT," 2010 4th International Conference on Bioinformatics and Biomedical Engineering, Chengdu, China, 2010, pp. 1-4.

[4] Y. Meng, L. Shi, J. Wang and P. Shi, "Considering Data Uncertainty in Species Sensitivity Distribution for Ecological Risk Assessment of Chemicals," 2010 4th International Conference on Bioinformatics and Biomedical Engineering, Chengdu, China, 2010, pp. 1-4.

[5] R. N. Kesavan and K. Latha, "Predicting Dynamic Distribution of Data Deficient Species under Climate Change," 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), Vellore, India, 2020, pp. 1-4.