Ecological Indicators 81 (2017) 362–372

Contents lists available at ScienceDirect

Ecological Indicators

journal homepage: www.elsevier.com/locate/ecolind

Water eutrophication evaluation based on semi-supervised classification: A case study in Three Gorges Reservoir

Di Wu[a,b,1], Huyong Yan[a,b,1], Mingsheng Shang[a], Kun Shan[a], Guoyin Wang[a,*]

[a] Chongqing Key Laboratory of Big Data and Intelligent Computing, Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, Chongqing 400714, China

[b] University of Chinese Academy of Sciences, Beijing 100049, China

ARTICLE INFO

Keywords:
Semi-supervised classification
Eutrophication evaluation
Three Gorges Reservoir

ABSTRACT

Water eutrophication, which refers to the enrichment of nutrients to an aquatic environment, is one of the most challenging problems in water protection. Although many researchers have made attempts to solve the eutrophication problem, there is one issue that needs to be further discussed, i.e., how to establish a fast, low-cost, and accurate eutrophication evaluation model? For addressing this issue, this paper proposes a data-driven eutrophication evaluation model based on the semi-supervised classification. Concretely, a case study in Three Gorges Reservoir of China is carried out to demonstrate the validity of the proposed model. Experimental results clearly show that the proposed model has the advantages of high computational efficiency, high accuracy, and great ability of exploiting low-cost factors to assist or even replace high-cost factors in realizing the eutrophication evaluation. Moreover, we find that three low-cost factors, including *pH*, dissolved oxygen, and ammonia-nitrogen, are effective in achieving a better eutrophication evaluation for Three Gorges Reservoir based on the proposed model.

1. Introduction

Water is one of the most important resources for human survival and economic development (Li et al., 2016a; Li et al., 2015). Unfortunately, water, including groundwater and surface-water, is undergoing different degrees of deterioration all over the world because of the intensive human activity (Wu and Sun, 2016). For example, serious groundwater pollution in western China is induced by rapid urbanization and industrialization (Li, 2016; Li et al., 2016b), and the degraded water quality in the United States is caused by overfishing and increased aquaculture (Heisler et al., 2008). Water eutrophication, which refers to the enrichment of nutrients (nitrogen and phosphorus) to an aquatic environment, is one of the most challenging problems in water protection (Heisler et al., 2008). Similarly, due to the increased discharge of nutrients from industrialization, agricultural modernization, and urbanization, eutrophication is often reported and has attracted many attentions from both public and government (Du et al., 2011; Phillips et al., 2013). Nitrogen and phosphorus are necessary elements for plant growth. However, if they are input into water body more than necessary, the ecosystem will be changed, such as harmful algal blooms (HABs) and high levels of phytoplankton biomass, resulting in degradation of water quality (Qin et al., 2013). Eutrophication has posed a threat to the safety of resident drinking water, the natural ecological environment, and economic development (Heisler et al., 2008). For instance, cyanobacterial blooms resulted from eutrophication caused a crisis for the Wuxi drinking water in 2007 (Qin et al., 2010). Thus, we must adopt proper measures to control eutrophication (Yang et al., 2008).

Obviously, the premise of controlling the water eutrophication is to establish the appropriate methods or models to evaluate the trophic status of water body. So far, many researchers have made attempts to address this issue. In the environmental and ecological fields, there are many models, including Carlson trophic state index (Carlson, 1977), modified Carlson trophic state index (Aizaki, 1981), trophic state (Vollenweider et al., 1998), comprehensive nutrition state index (Xu et al., 2012), phytoplankton trophic index (Phillips et al., 2013), species diversity index (Spatharis and Tsirtsis, 2013), integrated methodology (Wu et al., 2013), eutrophication index (Fertig et al., 2014), etc. In the field of informatics, neural network (Kuo et al., 2007; Melesse et al., 2016; Singh et al., 2012; Yang et al., 2015), genetic algorithm (Song et al., 2012), fuzzy set theory (Giusti et al., 2011), support vector machine (SVM) (Huo et al., 2014), and rough set theory (Yan et al., 2016b)

* Corresponding author.
E-mail addresses: wudi@cigit.ac.cn, wanggy@ieee.org (G. Wang).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.ecolind.2017.06.004
Received 31 January 2017; Received in revised form 30 May 2017; Accepted 4 June 2017

have been used in eutrophication evaluation. However, these methods or models still are not satisfactory in some cases because the eutrophication evaluation is nonlinear, multi-factors influenced, and complex in water ecological system (Ding and Wang, 2013). Additionally, due to the rapid development of automatic monitoring techniques (Dong et al., 2015), these methods or models are facing some new challenges as bellows:

- Due to the long time interval of traditional manual sampling, the collected monitoring data is small in most cases. Under such situation, the most existed eutrophication evaluation models were designed without considering the data processing ability. Thanks to the rapid development of automatic monitoring techniques, we can continuously collect massive monitoring data in 24 h. The massive monitoring data can effectively and accurately reflect the real conditions of water quality in time (Arienzo et al., 2015). Unfortunately, the explosive growth of monitoring data has brought some new challenges to the existed eutrophication evaluation models. For instance, How to effectively process the monitoring data with TB level is an important problem for the existed eutrophication evaluation methods or models.
- For most existed eutrophication evaluation models, total nitrogen (*TN*), total phosphorus (*TP*), Chlorophyll a (*Chl-a*), Secchi depth (*SD*), and Permanganate index (COD_{Mn}) are the key factors (Qin et al., 2013). Besides, some other factors, such as water temperature (*T*), dissolved oxygen (*DO*), *pH*, conductivity (*Cond.*), ammonia-nitrogen (NH_3-*N*), suspended solid (*SS*), position of sampling site (*site*), and season of sampling (*season*), are also connected with eutrophication (Giusti et al., 2011; Wu et al., 2013). However, the costs of collecting these factors are different because of their different monitoring principles. For example, *T, DO, pH, Cond.,* NH_3-*N*, and *SS*, can be collected easily by using the online sensors, while *TN, TP*, and COD_{Mn} are relatively difficult to collect because they demand the complicated pretreatment processes (A Xylem Brand, 2017). Thus, how to exploit the low-cost factors to assist or even replace the high-cost factors to realize the eutrophication evaluate is worth studying.
- In general, the existed eutrophication evaluation models rarely consider the ability of processing incomplete information. However, situation of missing key data happens frequently in the eutrophication evaluation because of the reasons of laboratory errors, instrument malfunctions, and even human errors. Therefore, it is necessary to take into account the ability of processing incomplete information for an eutrophication evaluation.

Classification, which relies heavily on the training instances with class labels, is an active research issue in data mining and machine learning communities (Cococcioni et al., 2012; Luo et al., 2015a; Su et al., 2009). Nevertheless, due to the technical support from experts as well as long time consumption of manual process for data labeling, it is difficult to obtain sufficient labeled data for supporting the classification tasks. Having a multitude of unlabeled data and few labeled ones is a common phenomenon in many practical applications (Zhou and Li, 2010). A successful and special methodology to tackle this problem is semi-supervised classification (SSC). SSC is highly effective in alleviating the shortage of labeled instances in classification tasks by exploiting the abundant unlabeled data (Triguero et al., 2015b). Thus, SSC has been widely and frequently used in several areas, including fault diagnosis, remote sensing monitoring, face recognition, *etc* (Schwenker and Trentin, 2014; Zhu, 2008). However, until now SSC has not been widely applied in eutrophication, and only few researchers have reported some preliminary studies on SSC applications in the fields of water supply (Herrera et al., 2010) and water quality retrieving (Wang et al., 2011).

Note that eutrophication evaluation actually is a classification task and the three challenges discussed above can be conquered by SSC. Thus, this paper innovatively proposes to use the SSC to achieve a powerful eutrophication evaluation model to overcome the three challenges discussed above. To the best of our knowledge, this paper is the first one to analyze eutrophication based on SSC and is different from the traditional environmental and ecological eutrophication analysis because it is completely data-driven. In order to illustrate the principle and usefulness of our proposal, a case study in Three Gorges Reservoir (TGR) of China is conducted in this paper. The experimental results clearly validate the fact that our proposal is a promising alternative method to achieve eutrophication evaluation with strong generalization ability.

The remainder of this paper is organized as follows: Section 2 introduces the materials and methods. Section 3 presents the proposed model. Section 4 provides and discusses the experimental results. Finally, Section 5 concludes this paper.

2. Materials and methods

2.1. Materials

2.1.1. Study area

TGR, which is created by the Three-Gorge Dam, sites at 29°16′–31°25′N, 106°–111°10′E, and has a surface area of 1080 km^2 (Zeng et al., 2006). The eutrophication in TGR has become the main environmental problem and attracts more and more attention from worldwide (Yan et al., 2016b). As shown in Fig. 1, the water level of TGR experiences two phases throughout the whole year because of flood management and water supply (Changjiang Maritime Safety Administration, 2017). In the ascending phase, the water level gradually increases from 145 *m* to 175 *m* and the flow speed becomes slow. The slow flow reduces the water exchange between the mainstream and the tributaries, resulting in the deposition of nutrients. Consequently, the eutrophication and even the HABs appeared in some tributaries (Yang et al., 2010). In the descending phase, the flow speed becomes fast, which intensifies the water exchange between the mainstream and the tributaries. As a result, water stability is reduced and the concentration of suspended silt is increased, which induces the frequent changes of water trophic state (Yang et al., 2010). Hence, the eutrophication evaluation of TGR should be high frequent for a better understanding on the water trophic state of TGR. Moreover, under such special hydrological conditions, the eutrophication evaluation of TGR also faces the three challenges discussed in Section 1. Therefore, we selected five typical tributaries in TGR as the study areas to validate the usefulness and effectiveness of our proposal, as shown in Fig. 2.

2.1.2. Field data

Data were obtained from several sampling sections located at five typical tributaries of TGR. The distributions of the tributaries are illustrated in Fig. 2. Sampling sections were visited during the period from

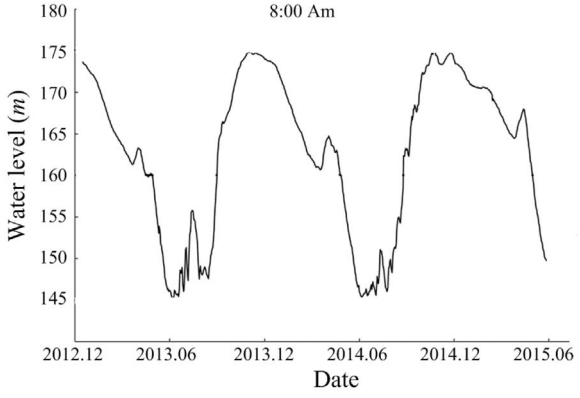

Fig. 1. The water level of TGR during two phases.

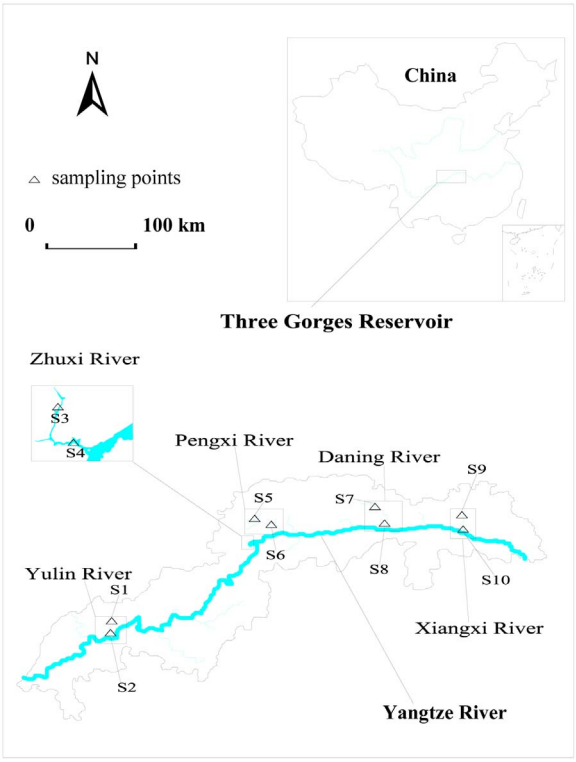

Fig. 2. The sampling sites of five typical tributaries in TGR.

Table 1
The analytical methods and standards of these samples.

Indicators	Units	Analytical methods	Standards and norms
Chla	mg/L	Spectrophotometry	SL88–2012
SD	m	Secchi disk method	/
TN	mg/L	Flow injection analysis and N-(1-naphthyl)ethylene diamine dihydrochloride spectrophotometry	HJ 668–2013
TP	mg/L	Flow injection analysis and ammonium molybdate spectrophotometry	HJ 671–2013
COD_{Mn}	mg/L	Acidic potassium permanganate method	GB/T11892–1989
T	°C	Thermometer	GB/T 13195–1991
Cond.	μs/m	Conductivity meter	/
pH	/	Glass electrode	GB/T6920–1986
DO	mg/L	Electrochemical probe method	HJ 506–2009
SS	mg/L	Gravimetric method	GB/T 11901-1989
NH_3-*N*	mg/L	Flow injection analysis and Salicylic acid spectrophotometry	HJ 666–2013

January 2007 to December 2015. Sampling frequency was once a month. To evaluate the eutrophication status, biological and chemical samples were taken at each sampling section. The analytical methods and standards shown in Table 1 were followed to ensure the analytical reliability of the samples. Besides, duplicates were also executed during the analyses for quality assurance and quality control.

2.2. Methods

2.2.1. Definition of SSC

SSC has been deeply investigated by prior researchers owing to its effectiveness in alleviating the shortage of labeled instances. Various SSC approaches can be generally classified into four categories, i.e., generative methods, graph-based methods, semi-supervised support vector machines, and self-labeled methods (Dong et al., 2016). In SSC, a sample of data is described with a *d*-dimensional vector of attributes plus on class label as:

$$X_i = (X_i^1, X_i^2, ..., X_i^d, \omega) \quad (1)$$

where X_i is the *i*th instance of all the data, *d* is the *d*th feature of X_i, and ω indicates X_i belongs to a class ω. *L* is labeled set with ω known and consists of N_L instances. *U* is unlabeled set with ω unknown and consists of N_U instances. Note that a typical SSC problem is $N_U > > N_L$. Set $L \cup U$ forms the training set T_R. In addition, there is a testing set T_S when new unseen instances have the same characteristics as X_i with ω unknown. The purpose of SSC is to learn a better classifier *C* by using T_R instead of only using *L* to predict the labels for unlabeled data (transductive, *U*) or new unseen data (inductive, T_S) (Triguero et al., 2015a).

2.2.2. Used SSC algorithm

Although there are many SSC algorithms, one of the most representative SSC algorithms, i.e., *Co-forest* (Li and Zhou, 2007), is selected to exemplify the principle of our proposal in this paper because of its excellent performance in solving SSC problems. Note that any kind of SSC algorithms, such as propagable graph based algorithm (Ni et al., 2012) and synthetic data based algorithm (Triguero et al., 2015a), can be used as the SSC algorithm in our proposal.

Co-forest is one of the self-labeled methods of SSC. It integrates a well-known ensemble method named *Random Forest* (Breiman, 2001) to estimate the labeling confidence of unlabeled instances and easily produce the final hypothesis. *Co-forest* also requires neither that the data are described by sufficient and redundant attribute subsets nor special learning algorithms which frequently employ time-consuming cross validation in learning. The *Random Forest* is the base classifier in *Co-forest*. The pseudo code of *Co-forest* is presented in Table 2, where H_i is the ensemble of *N* random trees except for *i*th random tree, the function $SubSampled\left(U, \frac{\hat{e}_{i,t-1}W_{i,t-1}}{\hat{e}_{i,t}}\right)$ means randomly removing $|U| - \frac{\hat{e}_{i,t-1}W_{i,t-1}}{\hat{e}_{i,t}}$number of data from *U*, the function $EstimateError(H_i, L)$ means estimating the classification error rate of H_i by using *L*, and the function $Confidence(H_i, x_u)$ means estimating the confidence level for x_u by using H_i.

2.2.3. Used standard eutrophication evaluation criterion

In this paper, the Technical Guideline for Water Environmental Quality Assessment of Three Gorges Reservoir (TGWEQA-TGR) (Ministry of Environmental Protection of the People's Republic of China, 2010) is selected as the standard eutrophication evaluation criterion because it is designed for TGR exclusively. The formulas (2)–(7) and Table 3 specify the evaluation method of TGWEQA-TGR. In Table 3, the data were classified into five ranks. Rank 1 is oligotrophic, Rank 2 is mesotrophic, Rank 3 is light eutrophic, Rank 4 is medium eutrophic, and Rank 5 is heavy eutrophic.

$$TLI(chla) = 10\left[2.46 + \frac{\ln(\rho_{chla})}{\ln 2.5}\right] \quad (2)$$

$$TLI(TN) = 10\left[2.46 + \frac{1.6316 + 4.3067\ln(\rho_{TN})}{\ln 2.5}\right] \quad (3)$$

$$TLI(TP) = 10\left[2.46 + \frac{10.2862 + 2.8691\ln(\rho_{TP})}{\ln 2.5}\right] \quad (4)$$

$$TLI(SD) = 10\left[2.46 + \frac{2.6027 - 7.6079\ln(\rho_{SD})}{\ln 2.5}\right] \quad (5)$$

$$TLI(COD_{Mn}) = 10\left[2.46 - \frac{0.7204 - 4.1230\ln(\rho_{COD_{Mn}})}{\ln 2.5}\right] \quad (6)$$

where ρ_{chla}, ρ_{TN}, ρ_{TP}, ρ_{SD} and $\rho_{COD_{Mn}}$ represent the concentrations of *Chl-a, TN, TP, SD*, and COD_{Mn}, respectively.

Table 2
Pseudo code of Co-forest.

Input: The labeled set *L*, the unlabeled set *U*, the confidence threshold θ, the number of random trees *N*.

Process:

Construct a random forest consisting *N* random trees.

For $i \in \{1,\ldots,N\}$ do

$\hat{e}_{i,0} \leftarrow 0.5$

$W_{i,0} \leftarrow 0$

End for

$t \leftarrow 0$

Repeat until none of the trees in Random Forest changes

$t \leftarrow t+1$

For $i \in \{1,\ldots,N\}$ do

$\hat{e}_{i,t} \leftarrow EstimateError(H_i, L)$

$L'_{i,t} \leftarrow \varnothing$

If ($\hat{e}_{i,t} < \hat{e}_{i,t-1}$)

$U'_{i,t} \leftarrow SubSampled\left(U, \frac{\hat{e}_{i,t-1} W_{i,t-1}}{\hat{e}_{i,t}}\right)$

For each $x_u \in U'_{i,t}$ do

If $(Confidence(H_i, x_u) > \theta)$

$L'_{i,t} \leftarrow L'_{i,t} \cup \{(x_u, H_i(x_u))\}$

$W_{i,t} \leftarrow W_{i,t} + Confidence(H_i, x_u)$

End for

For $i \in \{1,\ldots,N\}$ do

If $(e_{i,t} W_{i,t} < e_{i,t-1} W_{i,t-1})$

$h_i \leftarrow \text{LearnRandomTree}(L \cup L'_{i,t})$

End for

End of Repeat

Output: $H^*(x) \leftarrow \arg\max_{y \in label} \sum_{i: h_i(x)=y} 1$

Table 3
The evaluation grades of eutrophication for TGR.

Rank	*TLI(Σ)*
1	TLI(*Σ*) ≤ 30
2	30 < *TLI(Σ)* ≤ 50
3	50 < *TLI(Σ)* ≤ 60
4	60 < *TLI(Σ)* ≤ 70
5	*TLI(Σ)* > 70

$$TLI(\Sigma) = \sum_{j=1}^{5} W_j \cdot TLI(j) \tag{7}$$

where $TLI(\Sigma)$ represents the comprehensive trophic level index, $TLI(j)$ is the trophic level index of each indicator (*Chl-a, TN, TP, SD*, and COD_{Mn}), W_j denotes the weight of each indicator (*Chl-a, TN, TP, SD*, and COD_{Mn}). Among them, $W_{Chl\text{-}a} = 0.5996$, $W_{TN} = 0.0718$, $W_{TP} = 0.1370$, $W_{SD} = 0.0075$, $W_{COD_{Mn}} = 0.1840$.

3. The proposed model

In this section, the proposed model is described, in which an SSC algorithm is exploited to achieve a powerful eutrophication evaluation model. Fig. 3 depicts the flowchart of the proposed model. First, a small part of data is labeled according to one of the standard evaluation criteria. Next, the labeled data with labels are used as *L* and the other data without labels are used as *U*. Specifically, the data in *U* have fewer factors than that in *L*. Meanwhile, some other factors absented in the standard evaluation criterion can also be added into *L* and *U* if they are beneficial for the learning of evaluation model. The *L* and *U* form the T_R. Then, an SSC algorithm is exploited to keep learning an eutrophication evaluation model on T_R until the stopping criteria are satisfied. Finally, the learned evaluation model is employed to evaluate the trophic state for new unseen data (inductive, T_S). It is noticed that the data in *U* (transductive, *U*) also have been labeled during the learning process.

In order to explain the principle of the proposed model more clearly, a case is illustrated in Fig. 4. A small number of data with factors of *TN, TP, Chl-a, SD*, and COD_{Mn} are labeled according to TGWEQA-TGR first and then are used as *L,* as shown in Fig. 4(a). Meanwhile, a large number of data without the high-cost factor of COD_{Mn} are used as *U*. Next, both *L* and *U* are employed to learn an evaluation model based on an SSC algorithm. Finally, the new unseen data without the high-cost factor of COD_{Mn} can been labeled or evaluated accurately based on the learned evaluation model. Similarly, if some other low-cost factors absented in the TGWEQA-TGR, such as *DO*, *pH*, have assistance in learning the evaluation model, they can also be added into the whole processes to improve the performance of evaluation model, as shown in Fig. 4(b).

Obviously, the proposed model has the ability of using low-cost factors to assist or replace high-cost factors to evaluate eutrophication. In addition, it also has an excellent performance in calculating speed and accuracy if it integrates an efficient SSC algorithm. Thus, we expect that the proposed model can overcome all the challengers discussed in Section 1 and be a promising tool for practical eutrophication evaluation.

4. Experimental results and discussions

4.1. Dataset

A total of 587 pieces of data are collected in the experiments. Table 4 lists some representative data with 11 factors (condition attributes) and one Rank (decision attribute). The ranks of data are labeled according to TGWEQA-TGR. In order to analyze the correlations between Rank and each factor, we perform correlation analysis to conduct the correlation coefficient. Meanwhile, the correlations between one of the three factors (*TP, TN,* COD_{Mn}) and each factor are also analyzed respectively within the condition attributes. The results are shown in Table 5. According to Table 5, we find that the correlations between these parameters are statistically significant at a significance level $\alpha = 0.05$.

4.2. Cross validation analysis

In the experiments, we use a 5-fold cross-validation strategy to

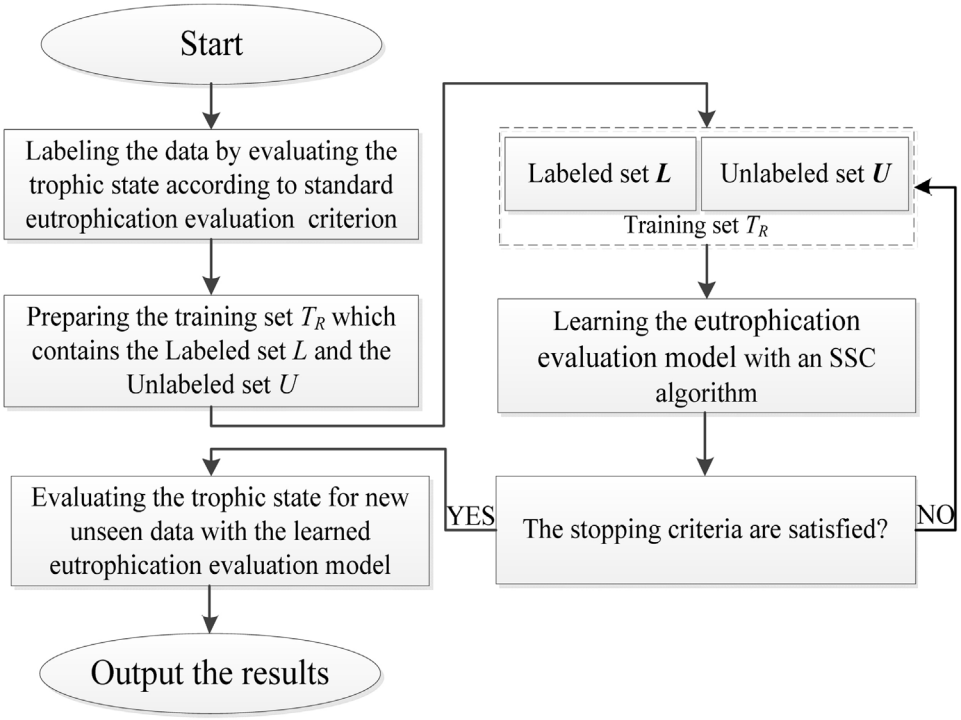

Fig. 3. Flowchart of the proposed model for eutrophication evaluation.

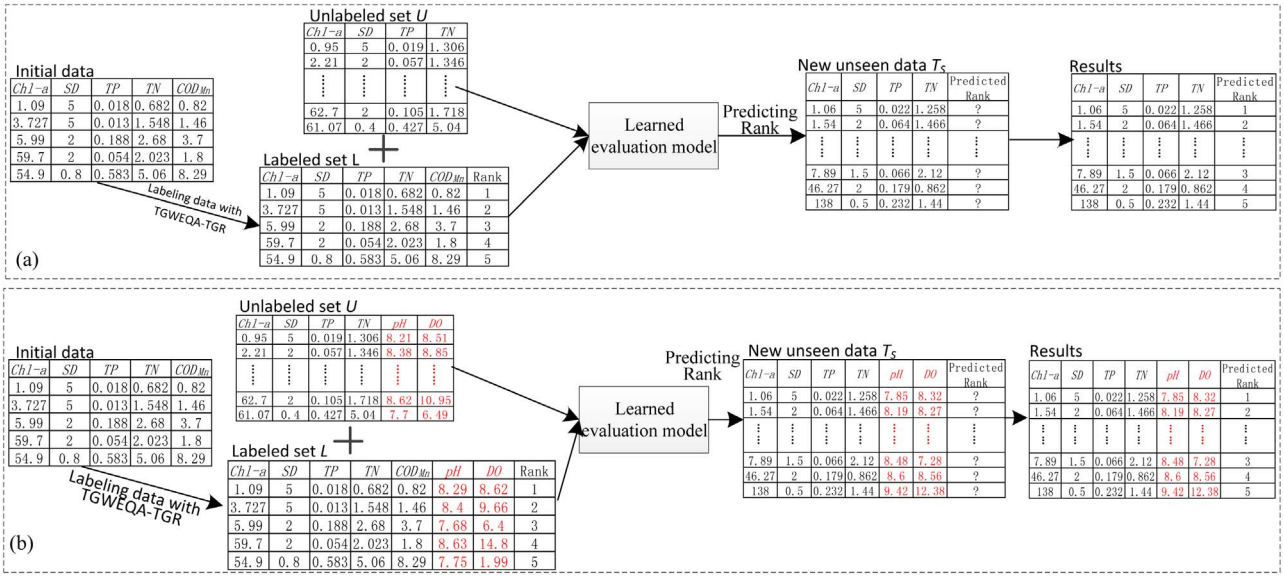

Fig. 4. A case of eutrophication evaluation based on the proposed model.

Table 4
Some representative data.

Record	*Chl-a*	*SD*	*TP*	*TN*	COD_{Mn}	*T*	*Cond.*	*pH*	*DO*	*SS*	*NH3-N*	Rank
1	1.06	5	0.022	1.258	1.11	23.8	286.5	7.85	8.32	4	0.081	1
2	1.54	2	0.064	1.466	1.37	20.1	285.7	8.19	8.27	5.5	0.143	2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
585	7.89	1.5	0.066	2.12	3.1	18.3	484	8.48	7.28	26	0.37	3
586	46.27	2	0.179	0.862	3.4	22.5	321.8	8.6	8.56	21	0.314	4
587	138	0.5	0.232	1.44	6	25	394	9.42	12.38	19	0.413	5

Table 5
The results of correlation analysis.

Parameters	*Chl-a*	*SD*	*TP*	*TN*	COD_{Mn}	*T*	*Cond.*	*pH*	*DO*	*SS*	*NH3-N*
Rank	0.612	−0.456	0.713	0.646	0.751	0.090	0.218	0.103	0.086	0.275	0.542
TP	0.472	−0.295	1.000	0.811	0.678	−0.045	0.313	−0.105	−0.104	0.152	0.746
TN	0.495	−0.307	0.811	1.000	0.667	−0.031	0.310	−0.131	−0.153	0.162	0.771
COD_{Mn}	0.618	−0.457	0.678	0.667	1.000	0.152	0.314	−0.029	−0.094	0.229	0.603

determine the final performances of the proposed model. First, the dataset is randomly split into five folds, each of which contains 20% of the instances. Then, four folds are selected to use as the T_R and the remaining one forms the T_S. After that, the T_R is divided into labeled part L and unlabeled part U by using a random stratified selection, which means that the selected number of instances for each rank is proportional to the number of them in the T_R. Besides, we will ensure that at least one representative instance of each rank is selected in the L. The value of L/T_R is called as the ratio of labeled data. Thus, each data is divided into three parts: L, U, and T_S (L and U form T_R). Subsequently, the proposed model is trained on the T_R, and then is tested on the U (transductive) and the T_S (inductive). Since there are some random operations during the stratified selection, we will repeat the training and testing processes for three times. The above steps will be executed five times to ensure that each fold can serve as the T_S once.

In order to quantitatively evaluate the performance of the proposed model, two quantities of *Accuracy* and *Standard Deviation* are computed respectively. The computing formulas are shown as below:

$$Accuracy_k = \frac{1}{n_k}\sum_{i=1}^{n_k}\psi(\omega, f(X_i)),\ \psi(\omega, f(X_i)) = \begin{cases}1, & if\ \omega = f(X_i)\\ 0, & else\end{cases} \tag{8}$$

$$Accuracy = \frac{1}{5}\sum_{k=1}^{5} Accuracy_k \tag{9}$$

$$Standard\ Deviation = \sqrt{\frac{1}{5}\sum_{k=1}^{5}(Accuracy_k - Accuracy)^2} \tag{10}$$

where $k \in \{1,2,\ldots,5\}$, X_i represents one piece of data, ω indicates the real rank value of X_i, $f(X_i)$ is the predicted rank value of X_i based on the proposed model, and n_k is the number of X_i for each evaluation. *Accuracy* represents the predictive ability and *Standard Deviation* represents the robustness of the proposed model.

4.3. Experiments and results

First, the behaviors of the proposed model are analyzed under the conditions of missing some factors and different ratio of labeled data. The situations of missing some factors consist of missing *Chl-a*, *SD*, *TP*, *TN*, and COD_{Mn} respectively, i.e., the case illustrated in Fig. 4(a). The ratio of labeled data is ranged from 10% to 50%. The proposed model with *Co-forest* is compared with the *Random Forest*. The proposed model represents the side with an SSC idea and *Random Forest* which is the base classifier of *Co-forest* represents the side without an SSC idea. The

Table 6
The results of "*Accuracy ± Standard Deviation*" on transductive.

Missing factor	*Random Forest*					The proposed model				
	Ratio of labeled data					Ratio of labeled data				
	10%	20%	30%	40%	50%	10%	20%	30%	40%	50%
No	81.98 ± 1.36	85.81 ± 3.71	87.88 ± 1.22	87.99 ± 1.63	88.56 ± 0.93	**82.67 ± 2.9**	**86.85 ± 1.56**	87.84 ± 0.73	**89.84 ± 0.7**	**89.95 ± 1.5**
Chl-a	66.46 ± 2.25	69.93 ± 2.44	69.23 ± 2.96	69.39 ± 1.07	70.18 ± 3.2	**69.16 ± 2.44**	69.91 ± 2.08	**70.02 ± 3.31**	**69.93 ± 0.81**	**70.66 ± 3**
SD	82.61 ± 5	83.56 ± 3.56	86.42 ± 1.52	88.43 ± 1.37	89.64 ± 1.06	**82.93 ± 3.16**	**85.7 ± 2.64**	**87.72 ± 0.7**	**88.5 ± 1.59**	**90.73 ± 1.19**
TP	81.88 ± 4.31	84.57 ± 1.48	84.86 ± 2.98	87.3 ± 1.07	88.78 ± 2.41	**83.16 ± 2.23**	**86.03 ± 1.94**	**86.88 ± 1.67**	**88.27 ± 1.75**	**89.24 ± 1.31**
TN	81.21 ± 2.19	84.6 ± 1.51	87.29 ± 3.18	87.66 ± 1.58	86.65 ± 1.8	**82.32 ± 1.3**	**86.57 ± 1.15**	**87.31 ± 0.75**	**89.6 ± 0.81**	**89.18 ± 1**
COD_{Mn}	79 ± 4.06	83.56 ± 3.68	84.66 ± 2.36	85.41 ± 2.36	88.97 ± 1.32	**81.04 ± 4.03**	**86.82 ± 1.54**	**86.87 ± 1.8**	**86.59 ± 1.25**	**89.97 ± 1.48**

Table 7
The results of "*Accuracy ± Standard Deviation*" on inductive.

Missing factor	*Random Forest*					The proposed model				
	Ratio of labeled data					Ratio of labeled data				
	10%	20%	30%	40%	50%	10%	20%	30%	40%	50%
No	80.23 ± 4.14	81.09 ± 2.7	87.39 ± 6.03	86.88 ± 1.95	89.44 ± 2.05	**80.92 ± 2.81**	**84.05 ± 3.05**	86.47 ± 5.11	**88.02 ± 1.83**	89.04 ± 2.38
Chl-a	65.6 ± 6.63	70.69 ± 6.36	68.31 ± 2.91	68.83 ± 2.19	70.52 ± 4.03	**69.98 ± 6.24**	**70.86 ± 4.09**	**68.59 ± 2.86**	**69.29 ± 3.23**	69.91 ± 4.53
SD	81.78 ± 5.09	82.11 ± 5.93	86.54 ± 5.8	87.57 ± 2.48	88.59 ± 2.79	**82.17 ± 4.16**	**84.85 ± 3.12**	**87.62 ± 1.91**	**89.22 ± 1.8**	**89.78 ± 2.92**
TP	83.32 ± 6.97	83.14 ± 3.69	86.37 ± 1.61	87.06 ± 3.93	88.58 ± 1.71	**83.93 ± 2.81**	**84.95 ± 2.87**	**87.96 ± 2.03**	**87.06 ± 3.47**	88.57 ± 3.85
TN	80.06 ± 4.67	83.3 ± 2.41	86.21 ± 5.15	87.56 ± 0.8	88.92 ± 4.64	**83.42 ± 2.84**	**86.14 ± 1.74**	**87.17 ± 2.55**	**89.49 ± 2.63**	**89.55 ± 3.16**
COD_{Mn}	80.24 ± 3.71	82.46 ± 1.97	86.03 ± 3.31	86.37 ± 2.09	89.1 ± 3.98	**81.48 ± 3.43**	**85.41 ± 1.49**	**87.33 ± 3.64**	**87.96 ± 3.1**	**89.21 ± 2.88**

Table 8
The results of Wilcoxon signed-ranks test with a significance level $\alpha = 0.1$ in the aspect of missing different factors: the proposed model VS. *Random Forest*.

Missing factor	Transductive						Inductive					
	Accuracy			*Standard Deviation*			*Accuracy*			*Standard Deviation*		
	$R+$	$R-$	*p*-value	$R+$	$R-$	*p*-value	$R+$	$R-$	*p*-value	$R+$	$R-$	*p*-value
No	14	1	0.0625	9	6	0.4063	11	4	0.2188	10	5	0.3125
Chl-a	14	1	0.0625	10	5	0.3125	11	4	0.2188	8	7	0.5000
SD	15	0	**0.0313**	12	3	0.1563	15	0	**0.0313**	14	1	**0.0625**
TP	15	0	**0.0313**	12	3	0.1563	13	2	**0.0938**	10	5	0.3125
TN	15	0	**0.0313**	15	9	**0.0313**	15	0	**0.0313**	12	3	0.1563
COD_{Mn}	15	0	**0.0313**	13	2	**0.0938**	15	0	**0.0313**	9	6	0.4063

Table 9
The results of Wilcoxon signed-ranks test with a significance level $\alpha = 0.1$ in the aspect of different ratio of labeled data: the proposed model VS. *Random Forest*.

Ratio of labeled data	Transductive						Inductive					
	Accuracy			*Standard Deviation*			*Accuracy*			*Standard Deviation*		
	$R+$	$R-$	*p*-value	$R+$	$R-$	*p*-value	$R+$	$R-$	*p*-value	$R+$	$R-$	*p*-value
10%	21	0	0.0156	15	6	0.2188	21	0	0.0156	21	0	**0.0156**
20%	20	1	**0.0313**	18	3	**0.0781**	21	0	**0.0156**	18	3	**0.0781**
30%	19	2	**0.0469**	20	1	**0.0313**	19	2	**0.0469**	16	5	0.1563
40%	21	0	**0.0156**	17	4	0.1094	21	0	**0.0156**	6	15	0.8438
50%	21	0	**0.0156**	14	7	0.2813	13	8	0.3438	7	14	0.6563

recommended values of parameters for *Co-forest* and *Random Forest* are chosen according to the original paper, i.e., confidence threshold θ is 0.75 and number of random trees N is 6. The obtained results of *Accuracy* and *Standard Deviation* on transductive and inductive settings are shown in Tables 6 and 7 respectively. Where the proposed model performs better than *Random Forest* are highlighted in bold.

According to Tables 6 and 7, there are two observations that we can conclude preliminarily: (1) The models with one missing factor have similar performances as that without missing factor except for the case of missing *Chl-a*. The reason can be ascribed to the maximal weight of *Chl-a* in TGWEQA-TGR. (2) The proposed model generally gives better results than *Random Forest*.

In order to further analyze the significant differences between *Random Forest* and the proposed model, we perform non-parametric pairwise comparison procedures to conduct the statistical analysis. Concretely, the Wilcoxon signed-ranks test (Wilcoxon, 1945) is applied to realize it. Table 8 records the statistical results in the aspect of missing different factors. Table 9 records the statistical results in the aspect of different ratio of labeled data. In Tables 8 and 9, three parameters show the achieved rankings $R+$ and $R-$ values and their associate *p*-value. We have checked whether the proposed model outperforms *Random Forest* under the condition of significance level $\alpha = 0.1$ and the accepted hypotheses are highlighted in bold.

According to Tables 8 and 9, two observations are concluded: (1) In either situation, the statistical results accept the hypothesis that the proposed model has higher *Accuracy* than *Random Forest* in most cases. Although there are three cases that the hypothesis is not accepted, the proposed model also achieves the higher $R+$ rankings of *Accuracy* than *Random Forest*, which reflects that the proposed model performs slightly better. (2) In the aspect of missing different factors, the proposed model achieves significantly or slightly better *Standard Deviation* than *Random Forest*. Meanwhile, in the aspect of different ratio of labeled data, the proposed model also has significantly or slightly better *Standard Deviation* than *Random Forest* in most cases. The two cases that the proposed model has lower $R+$ rankings of *Standard Deviation* than *Random Forest* may be explained by one major reason. That is the random selection operation during the learning process.

In summary, the experimental results demonstrate two conclusions: (1) The proposed model can realize the eutrophication evaluation under the condition of having a small amount of data without missing factor and a large amount of data with missing some factors. (2) The proposed model has a better performance than *Random Forest*, which means that SSC is a suitable tool to improve the performance of an eutrophication evaluation model trained only on a small amount of data without missing factor.

4.4. Discussions

Some reports indicate that eutrophication is not only connected with the five key factors (*Chl-a, SD, TN, TP, and* COD_{Mn}), but also has relationships with some other factors (*T, Cond., pH, DO, SS,* and NH_3-*N*) (Giusti et al., 2011; Wu et al., 2013). Due to the different monitoring principles, the costs of collecting these factors are different. For example, *T, Cond., pH, DO, SS,* and NH_3-*N* can be collected easily by using the online sensors, while *TN, TP*, and COD_{Mn} are relatively difficult to collect because they demand the complicated pretreatment processes. Thus, we will research whether we can exploit the low-cost factors to assist the eutrophication evaluation under the condition of missing some high-cost factors, i.e., the case illustrated in Fig. 4(b).

First, we analyze the behaviors of the proposed model with missing three high-cost factors (*TN, TP,* and COD_{Mn}) and the results are regarded as the baseline for later comparisons. Next, one of the low-cost factors (*T*, *Cond.*, *pH*, *DO*, *SS*, and NH_3-*N*) is respectively added into the experiments to test whether it has assistance on the eutrophication evaluation. The results are recorded in Table 10, where the adding has assistance are highlighted in bold.

According to Table 10, we can preliminarily find that the performances of respective adding of *pH*, *DO*, and NH_3-*N* are better than the baseline. Furthermore, the performances of adding NH_3-*N* are best. Subsequently, other three kinds of adding factors are tested. They are adding *DO and* NH_3-*N*, adding *pH* and NH_3-*N*, and adding *DO*, *pH* and NH_3-*N*. The results are also recorded in Table 10. In order to further analyze the different performances achieved by adding different factors, we perform the Friedman test (Demšar, 2006) with the significance level $\alpha = 0.1$ to conduct the statistical analysis and the results are recorded in Table 11.

Table 10
The results of "*Accuracy* ± *Standard Deviation*" under the conditions of missing three factors of *TN, TP, and* COD_{Mn} and adding some other factors.

Adding factor	Transductive					Inductive				
	Ratio of labeled data					Ratio of labeled data				
	10%	20%	30%	40%	50%	10%	20%	30%	40%	50%
No	79.95 ± 2.8	81.95 ± 1.14	82.75 ± 1.1	83.52 ± 0.65	82.57 ± 1.85	80.22 ± 5.4	82.85 ± 4.02	83.3 ± 2.4	83.25 ± 2.7	81.71 ± 5.15
T	**81.68 ± 1.64**	81.58 ± 1.27	**83.36 ± 1.21**	83.39 ± 1.03	**83.32 ± 1.82**	**80.58 ± 2.97**	81.03 ± 2.95	82.34 ± 3.23	82.46 ± 3.74	**85.58 ± 2.07**
Cond.	77.64 ± 6.68	**83.07 ± 2.76**	**83.96 ± 1.79**	83.31 ± 0.96	**83.39 ± 2.51**	77.22 ± 4.5	82.56 ± 2.44	**84.09 ± 2.33**	82.56 ± 3.58	**83.36 ± 2.95**
pH	**81.79 ± 1.68**	**83.37 ± 2**	**82.96 ± 1.58**	**83.74 ± 1.5**	**85.14 ± 1.22**	**81.98 ± 6.27**	82.84 ± 3.83	82.05 ± 3.48	**85.57 ± 2.65**	**83.13 ± 3.58**
DO	**81.45 ± 2.17**	**82.29 ± 1.76**	**83.91 ± 0.76**	**84.87 ± 1.21**	**84.96 ± 2.75**	**81.66 ± 4.62**	82.11 ± 3.34	**84.73 ± 4.27**	83.02 ± 3.44	**83.71 ± 3.46**
SS	78.07 ± 4.96	81.68 ± 1.8	**84.05 ± 1.96**	**83.55 ± 1.25**	**84.96 ± 0.9**	79.62 ± 2.59	79.23 ± 5.72	82.91 ± 2.78	82.46 ± 1.55	**83.31 ± 3.32**
NH_3-*N*	**82.46 ± 1.47**	**84.86 ± 0.92**	**85.69 ± 1.82**	**86.98 ± 0.75**	**86.29 ± 1.39**	**83.01 ± 3.67**	**85.29 ± 0.9**	**85.06 ± 2.18**	**85.97 ± 1.73**	**86.42 ± 1.13**
DO/NH_3-*N*	**81.47 ± 2.61**	**83.38 ± 2.02**	**84.95 ± 1.4**	**84.63 ± 0.76**	**84.5 ± 1.62**	**81.26 ± 4.37**	**83.3 ± 1.78**	**85.18 ± 1.02**	**84.89 ± 4.46**	**87.16 ± 3.47**
pH/NH_3-*N*	**81.47 ± 2.61**	**83.71 ± 2.4**	**85.54 ± 2.51**	**85.06 ± 1.34**	**86.57 ± 3.05**	**80.75 ± 3.22**	82.23 ± 2.2	**85.01 ± 3.81**	**85.07 ± 0.55**	**86.6 ± 3.06**
DO/*pH*/NH_3-*N*	**80.5 ± 1.59**	**84.13 ± 1.22**	**86.08 ± 1.22**	**87.01 ± 1.87**	**87.53 ± 1.72**	**80.91 ± 5.38**	**84.49 ± 1.47**	**85.17 ± 3.4**	**85.68 ± 3.33**	**86.25 ± 3.63**

Table 11
The average values of rankings computed by Friedman test on Table 10.

Adding factor	Transductive		Inductive	
	Accuracy	*Standard Deviation*	*Accuracy*	*Standard Deviation*
No	2.20	7	4	3.6
T	3.20	6.8	3.2	6
Cond.	3.00	3.2	3.6	6
pH	5.80	5.4	5.2	3.4
DO	5.20	5.8	5.2	3.6
SS	3.80	4.6	2	6.4
NH_3-*N*	**9.20**	**8**	**9.2**	**8.8**
DO/NH_3-*N*	8.60	6.2	8	4.6
pH/NH_3-*N*	6.10	5.7	8.2	5.8
DO/*pH*/NH_3-*N*	7.90	2.3	6.4	6.8

According to Table 11, the Friedman test shows that there are significant differences in the *Accuracy* and *Standard Deviation*. Note that the higher value indicates the better performance. On transductive, we observe that any situation of adding factor has higher value of *Accuracy* than the baseline, while only the situation of adding NH_3-*N* has higher value of *Standard Deviation* than the baseline. On inductive, we observe that the four situations of adding factor (adding NH_3-*N*, adding *DO and* NH_3-*N*, adding *pH* and NH_3-*N*, and adding *DO*, *pH* and NH_3-*N*) have both higher values of *Accuracy* and *Standard Deviation* than the baseline. Furthermore, we can find that the situation of adding NH_3-*N* has the highest values of *Accuracy* and *Standard Deviation*, no matter on transductive or inductive.

In summary, the results of Friedman test demonstrate two conclusions under the condition of missing *TN, TP,* and COD_{Mn}: (1) Six situations of adding factor have assistance on the eutrophication evaluation. They are adding *pH*, adding *DO,* adding NH_3-*N*, adding *DO* and NH_3-*N*, adding *pH* and NH_3-*N*, and adding *DO*, *pH* and NH_3-*N* respectively. (2) The situation of adding NH_3-*N* has the most assistance among the six situations.

The above two conclusions may be explained from the aspects of ecology and environment: (1) NH_3-*N* not only is a part of the *TN* but also has the direct relationship with *TN*. Adding the NH_3-*N* can directly help the proposed model to learn more information about the eutrophication. Thus, the situation of adding NH_3-*N* has the most assistance for the proposed model to evaluate eutrophication. (2) *DO* and *pH* have indirect relationship with *Chl-a* because they will change according to *Chl-a*. Adding *DO* and *pH* actually input more information of *Chl-a* into the proposed model indirectly. Moreover, the *Chl-a* has the maximal weight in TGWEQA-TGR. Therefore, the situations of adding *DO* and *pH* have some assistance on the eutrophication evaluation. (3) Eutrophication actually is greatly affected by different seasons. The seasonal information is indicated by *T*. Obviously, we expect that adding *T* should have assistance on the eutrophication evaluation. However, the experimental results show that there is no significant improvement on the performance after adding *T*. The reason may be found from the data which are obtained from several sampling sections respectively located at five tributaries. The geographical difference may have eliminated the positive influence obtained from *T* for the eutrophication evaluation. (4) *Cond.* is a dependent variable of *SS*. Further, *SS* is more affected by the water flow speed. Obviously, the water flow speed in TGR is controlled by the water level. As a result, *Cond.* and *SS* are more related with water level. Consequently, the situations of adding *Cond.* and *SS* have no assistance on the eutrophication evaluation.

4.5. Differences between the proposed model and some previous related models

Recently, we have made extensive research on the eutrophication in

Table 12
Summarizes the properties of the different eutrophication evaluation models.

Model	Application scenarios	Requirements
STRRA	Analyzing the spatial and temporal relation with eutrophication/ Small data.	Pre-labeling according to a criterion without data missing/Data without randomness and fuzziness.
RSMCM	Data-driven eutrophication evaluation/Small data.	Pre-labeling according to a criterion without data missing.
RSPN	Data-driven eutrophication evaluation/Big data.	Pre-labeling according to a criterion without data missing.
Proposed model	Data-driven eutrophication evaluation/Big data.	Pre-labeling according to a criterion with data missing or needing some other factors without in the criterion.

TGR and realized several evaluation models based on different methods. They are spatial and temporal relation rule acquisition based model (STRRA) (Yan et al., 2016b), rough set and multidimensional cloud model based model (RSMCM) (Yan et al., 2017), and rough set and petri nets based model (RSPN) (Yan et al., 2016a). STRRA model achieves the spatial and temporal relation rule acquisition of eutrophication without any prior knowledge. Yet it fails to properly cope with the data with randomness and fuzziness, which is solved by introducing multidimensional cloud model into the eutrophication evaluation in RSMCM model. However, there is a problem in the above two models: the searching of knowledge is very time-consuming in the context of massive data. In this regard, RSPN realizes the fast knowledge inference by using the parallel processing mechanism of petri nets and the knowledge reduction ability of rough set. While the premise of establishing the above three models is that the data must be complete without any missing. As discussed in Sections 3 and 4, we know that the proposed model can not only accomplish the eutrophication evaluation under the condition of missing some high-cost factors, but also exploit some other low-cost factors to further improve its performance of eutrophication evaluation, which is the biggest difference between the proposed model and the three models of STRRA, RSMCM, and RSPN. Thus, we summarize the properties of these eutrophication evaluation models in Table 12.

4.6. Possible extensions

Although the proposed model has shown the promising prospect, there are several open issues should be considered. For example, are there some other low-cost factors that can be used to achieve our proposed mode? In addition, how to improve our proposed model's performance of eutrophication evaluation can be further studied. Lastly, it is unknown whether our proposed mode can be applied to other environmental fields. Therefore, in the future, we plan to test whether meteorological factors and geographic factors can be used to achieve our proposed mode at first. Subsequently, we will take advantage of ensemble technology (Rokach, 2010) to further improve our proposed model. Finally, we will extend our proposed model to the online water quality monitoring system first and then apply it in water resource management (Luo et al., 2013), flood management (Luo et al., 2015a,b), and water quality assessment (Li et al., 2012, 2011).

5. Conclusions

In this paper, we propose to use the SSC to achieve a powerful eutrophication evaluation. To the best of our knowledge, this paper is the first one to analyze eutrophication by using the SSC technology. The proposed model is different from the traditional environmental and ecological eutrophication analysis because it is completely data-driven. Concretely, a case study in TGR of China is conducted to demonstrate the validity of the proposed model. Experimental results clearly reflect the fact that the proposed model can achieve a high performance eutrophication evaluation model under the condition of having a small amount of data without missing factor and a large amount of data with missing some factors. Besides, we also demonstrate that the proposed model can exploit the low-cost factors (*pH*, *DO*, and NH_3-*N*) to assist the eutrophication evaluation under the condition of missing some high-cost factors (*TN*, *TP*, and COD_{Mn}). Thus, the proposed model has the advantages of high computational efficiency, high accuracy, and excellent ability of exploiting low-cost factors to assist or even replace high-cost factors to achieve eutrophication evaluation. We expect that the proposed model will have practical application in environmental protection sooner or later.

Acknowledgments

This work was supported by the support of the national key scientific and technological project of China (2014ZX07104-006), the National Natural Science Foundation of China (NSFC) (No.61272060), the Application Development Plan Project of Chongqing (cstc2014yykfC0053), and the Hundred Talents Program of CAS (No. Y500091BR1). The authors would like to thank the editor and anonymous referees for their excellent comments and valuable suggestions to improve the composition and content substantially.

References

A Xylem Brand, 2017. Homepage of YSI. Available online at https://www.ysi.com/. (Accessed 11 April, 2017).
Aizaki, M., 1981. Application of modified Carlson's trophic state index to Japanese lakes and its relationships to other parameters related to trophic state. Res. Rep. Natl. Inst. Environ. Stud. Jpn. 23, 13–31 (in Japanese).
Arienzo, A., Sobze, M.S., Wadoum, R.E., Losito, F., Colizzi, V., Antonini, G., 2015. Field application of the micro biological survey method for a simple and effective assessment of the microbiological quality of water sources in developing countries. Int. J. Environ. Res. Public Health 12, 10314–10328.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
Carlson, R.E., 1977. A trophic state index for lakes. Limnol. Oceanogr. 22, 361–369.
Changjiang Maritime Safety Administration, 2017. Water Level Announcement. Available online at http://www.cjmsa.gov.cn/9/368/2/39/312/, in Chinese. (Accessed 11 April, 2017).
Cococcioni, M., Lazzerini, B., Volpi, S.L., 2012. Robust diagnosis of rolling element bearings based on classification techniques. IEEE. Trans. Ind. Inform. 9, 2256–2263.
Demšar, J., 2006. Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. 7, 1–30.
Ding, H., Wang, D., 2013. The evaluation method of water eutrophication based on cloud model. Acta Sci. Circumst. 33, 251–257.
Dong, J., Wang, G., Yan, H., Xu, J., Zhang, X., 2015. A survey of smart water quality monitoring system. Environ. Sci. Pollut. Res. 22, 4893–4906.
Dong, A., Chung, F.l., 2016. S. Wang Semi-supervised classification method through oversampling and common hidden space. Inf. Sci. 349, 216–228.
Du, L.N., Li, Y., Chen, X.Y., Yang, J.X., 2011. Effect of eutrophication on molluscan community composition in the Lake Dianchi (China, Yunnan). Limnologica 41, 213–219.
Fertig, B., Kennish, M.J., Sakowicz, G.P., Reynolds, L.K., 2014. Mind the data gap:identifying and assessing drivers of changing eutrophication condition. Estuaries Coasts 37, 198–221.
Giusti, E., Marsilillibelli, S., Mattioli, S., 2011. A fuzzy quality index for the environmental assessment of a restored wetland. Water Sci. Technol. 63, 2061–2070.
Heisler, J., Glibert, P.M., Burkholder, J.M., Anderson, D.M., Cochlan, W., Dennison, W.C., Dortch, Q., Gobler, C.J., Heil, C.A., Humphries, E., 2008. Eutrophication and harmful algal blooms: a scientific consensus. Harmful Algae 8, 3–13.
Herrera, M., Canu, S., Karatzoglou, A., Izquierdo, J., Perez-Garcia, R., 2010. An approach to water supply clusters by semi-supervised learning. Proc. Int. Environ.Modell. Softw. Soc. (iEMSs).
Huo, A., Zhang, J., Qiao, C., Li, C., Xie, J., Wang, J., Zhang, X., 2014. Multispectral remote sensing inversion for city landscape water eutrophication based on Genetic Algorithm-Support Vector Machine. Water Qual. Res. J. Can. 49, 285–293.
Kuo, J.-T., Hsieh, M.-H., Lung, W.-S., She, N., 2007. Using artificial neural network for reservoir eutrophication prediction. Ecol. Modell. 200, 171–177.
Li, M., Zhou, Z.-H., 2007. Improve computer-aided diagnosis with machine learning

techniques using undiagnosed samples. IEEE. Trans. Syst. Man Cybern. A 37, 1088–1098.
Li, P.Y., Qian, H., Wu, J.H., 2011. Application of set pair analysis method based on entropy weight in groundwater quality assessment – a case study in Dongsheng city, Northwest China. E-J. Chem. 8, 851–858.
Li, P.Y., Wu, J., Hui, Q., 2012. Groundwater quality assessment based on rough sets attribute reduction and TOPSIS method in a semi-arid area. China Environ. Monit. Assess. 184, 4841–4854.
Li, P.Y., Qian, H., Howard, K.W.F., Wu, J., 2015. Building a new and sustainable Silk Road economic belt. Environ. Earth Sci. 74, 7267–7270.
Li, P.Y., Feng, W., Xue, C., Tian, R., Wang, S., 2016a. Spatiotemporal variability of contaminants in lake water and their risks to human health: a case study of the Shahu lake tourist area. Northwest China. Expo. Health. http://dx.doi.org/10.1007/s12403-016-0237-3.
Li, P.Y., Wu, J., Qian, H., Zhang, Y., Yang, N., Jing, L., Yu, P., 2016b. Hydrogeochemical characterization of groundwater in and around a wastewater irrigated forest in the southeastern edge of the Tengger desert, Northwest China. Expo. Health 8, 331–348.
Li, P.Y., 2016. Groundwater quality in western China: challenges and paths forward for groundwater quality research in Western China. Expo Health 8, 305–310.
Luo, P., He, B., Chaffe, P.L., Nover, D., Takara, K., Ma, M.R.R., 2013. Statistical analysis and estimation of annual suspended sediments of major rivers in Japan. Environ. Sci. Process. Impacts 15, 1052–1061.
Luo, F.J., Dong, Z.Y., Chen, G., Xu, Y., Meng, K., Chen, Y.Y., Wong, K.P., 2015a. Advanced pattern discovery-based fuzzy classification method for power system dynamic security assessment. IEEE Trans. Ind. Inform. 11, 416–426.
Luo, P., He, B., Takara, K., Xiong, Y.E., Nover, D., Duan, W., Fukushi, K., 2015b. Historical assessment of Chinese and Japanese flood management policies and implications for managing future floods. Environ. Sci. Policy 48, 265–277.
Melesse, A.M., Krishnaswamy, J., Zhang, K., 2016. Modeling coastal eutrophication at Florida Bay using neural networks. J. Coast. Res. 24, 190–196.
Ministry of Environmental Protection of the People's Republic of China, 2010. Technical Guideline for Water Environmental Quality Assessment of Three Gorges Reservoir. Available online at http://www.zhb.gov.cn/gkml/hbb/bgth/201012/t20101217_198815.htm, in Chinese. (Accessed 11 April, 2017).
Ni, B., Yan, S., Kassim, A.A., 2012. Learning a propagable graph for semisupervised learning: classification and regression. IEEE Trans Knowl. Data Eng. 24, 114–126.
Phillips, G., Lyche-Solheim, A., Skjelbred, B., Mischke, U., Drakare, S., Free, G., Järvinen, M., Hoyos, C.D., Morabito, G., Poikane, S., 2013. A phytoplankton trophic index to assess the status of lakes for the Water Framework Directive. Hydrobiologia 704, 75–95.
Qin, B., Zhu, G., Gao, G., Zhang, Y., Wei, L., Paerl, H.W., Carmichael, W.W., 2010. A drinking water crisis in lake taihu, China: linkage to climatic variability and lake management. Environ. Manage. 45, 105–112.
Qin, B.Q., Gao, G., Zhu, G.W., Zhang, Y.L., Song, Y.Z., Tang, X.M., Hai, X.U., Deng, J.M., 2013. Lake eutrophication and its ecosystem response. Sci.Bull 58, 961–970.
Rokach, L., 2010. Ensemble-based classifiers. Artif. Intell. Rev. 33, 1–39.
Schwenker, F., Trentin, E., 2014. Pattern classification and clustering:A review of partially supervised learning approaches. Pattern. Recogn. Lett. 37, 4–14.
Singh, K.P., Gupta, S., Singh, K.P., Gupta, S., 2012. Artificial intelligence based modeling for predicting the disinfection by-products in water. Chemom. Intell. Lab. Syst. 114, 122–131.
Song, K., Li, L., Tedesco, L.P., Li, S., Clercin, N.A., Hall, B.E., Li, Z., Shi, K., 2012. Hyperspectral determination of eutrophication for a water supply source via genetic algorithm-partial least squares (GA-PLS) modeling. Sci. Total Environ. 426, 220–232.
Spatharis, S., Tsirtsis, G., 2013. Zipf–Mandelbrot model behavior in marine eutrophication: two way fitting on field and simulated phytoplankton assemblages. Hydrobiologia 714, 191–199.
Su, Y., Shan, S., Chen, X., Gao, W., 2009. Hierarchical ensemble of global and local classifiers for face recognition. IEEE. Trans. Image Process. 18, 1885–1896.
Triguero, I., García, S., Herrera, F., 2015a. SEG-SSC: a framework based on synthetic examples generation for self-labeled semi-supervised classification. IEEE. Trans. Cybern. 45, 622–634.
Triguero, I., García, S., Herrera, F., 2015b. Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. Knowl. Inf. Syst. 42, 245–284.
Vollenweider, R.A., Giovanardi, F., Montanari, G., Rinaldi, A., 1998. Characterization of the trophic conditions of marine coastal waters with special reference to the NW Adriatic Sea: proposal for a trophic scale, turbidity and generalized water quality index. Environmetrics 9, 329–357.
Wang, X., Fu, L., Ma, L., 2011. Semi-supervised support vector regression model for remote sensing water quality retrieving. Chin. Geogr. Sci. 21, 57–64.
Wilcoxon, F., 1945. Individual comparisons by ranking methods. Biom. Bull. 1, 80–83.
Wu, J., Sun, Z., 2016. Evaluation of shallow groundwater contamination and associated human health risk in an alluvial plain impacted by agricultural and industrial activities, Mid-west China. Expo. Health. 8, 311–329.
Wu, Z., Yu, Z., Song, X., Yuan, Y., Cao, X., Liang, Y., 2013. Application of an integrated methodology for eutrophication assessment: a case study in the Bohai Sea. Chin. J. Oceanol. Limnol. 31, 1064–1078.
Xu, M.J., Yu, L., Zhao, Y.W., Li, M., 2012. The simulation of shallow reservoir eutrophication based on MIKE21: a case study of douhe reservoir in north China. Proc. Environ. Sci. 13, 1975–1988.
Yan, H.Y., Huang, Y., Wang, G., Zhang, X., Shang, M., Feng, L., Dong, J., Shan, K., Wu, D., Zhou, B., 2016a. Water eutrophication evaluation based on rough set and petri nets: a case study in Xiangxi-Rive, Three Gorges Reservoirr. Ecol. Indic. 69, 463–472.
Yan, H.Y., Zhang, X.R., Dong, J.H., Shang, M.S., Shan, K., Wu, D., Yuan, Y., Wang, X., Meng, H., Huang, Y., 2016b. Spatial and temporal relation rule acquisition of eutrophication in Da'ning River based on rough set theory. Ecol. Indic. 66, 180–189.
Yan, H.Y., Wu, D., Huang, Y., Wang, G., Shang, M., Xu, J., Shi, X., Shan, K., Zhou, B., Zhao, Y., 2017. Water eutrophication assessment based on rough set and multi-dimensional cloud model. Chemom. Intell. Lab. Syst. 164, 103–112.
Yang, X.E., Xiang, W.U., Hao, H.L., Zhen-Li, H.E., 2008. Mechanisms and assessment of water eutrophication. J. Zhejiang. Univ-SC. B 9, 197–209.
Yang, Z.J., Liu, D.F., Ji, D.B., Xiao, S.B., 2010. Influence of the impounding process of the Three Gorges Reservoir up to water level 172.5 m on water eutrophication in the Xiangxi Bay. Sci. China: Technol. Sci. 53, 1114–1125.
Yang, D., Chen, F., Zhou, Y., 2015. A novel eutrophication assessment models for aquaculture water area via artificial neural networks. J. Comput. Theor. Nanosci. 12, 357–360.
Zeng, H., Song, L., Yu, Z., Chen, H., 2006. Distribution of phytoplankton in the Three-Gorge Reservoir during rainy and dry seasons. Sci. Total Environ. 367, 999–1009.
Zhou, Z.H., Li, M., 2010. Semi-supervised learning by disagreement. Knowl. Inf. Syst. 24, 415–439.
Zhu, X., 2008. Semi-Supervised learning literature survey. Comput. Sci. 37, 63–77.