COMPARISON BETWEEN THE LINEAR MODEL AND K-NEAREST NEIGHBOR METHOD FOR PREDICTING MACROINVERTEBRATE ASSEMBLES IN A CITY RIVER IN BEIJING, CHINA

YANG, L.[1] – BAI, X.[2*] – HU, Y.[1]

[1]College of Geoscience and Surveying Engineering, China University of Mining & Technology 100083 Beijing, China

[2]Resource and Environment Branch, China National Institute of Standardization 100191 Beijing, China

**Corresponding author*
e-mail: baixue@cnis.gov.cn

(Received 15th Jul 2017; accepted 2nd Nov 2017)

Abstract. Benthic macroinvertebrates play an important role in materials and energy flow in river ecosystems. In this paper, we built models, a linear model and k-nearest neighbor method, for predicting biodiversity of macroinvertebrates in a city river using the data from Wenyu River. Both Shannon-Wiener index and Simpson index were considered for measuring the biodiversity of macroinvertebrates. The observed data of macroinvertebrates and 12 water quality indicators in Wenyu River, from 2010 to 2012, were applied in building and validating the predicted models. The results indicated that 1) The validity of the linear model was, though not perfect, better for predicting macroinvertebrates diversity using water quality indicators than k-nearest neighbor method in a city river; 2) Simpson index was more robust and accurate than the other biodiversity index to act as the variable of predicting benthic macroinvertebrates in a city river. There were 89.47% observations within the 99% confidence intervals. The developed predictive model was a useful tool for assessing river health, especially city river health, without taking into account the abundances of invertebrates.

Keywords: *macroinvertebrates, biodiversity, predicting models, city river, Beijing*

Introduction

Rivers are suffering biodiversity loss, water quality deterioration, hydrological changes, and channelization etc. (Davies et al., 2010; Pan et al., 2012). River restoration has become one of the important water environmental management problems. Benthic macroinvertebrates are proved to be valuable in conservation and ecological restoration of river ecosystems (Heino et al., 2003; Bae et al., 2005). Because of their confinement to the bottom, limited movement abilities and the long-life cycles, benthic macroinvertebrates are considered to be appropriate indicators for the evaluation of environments' long-term changes (Barbour et al., 1999; Timm and Mols, 2012; Pan et al., 2012; Hejazi et al., 2017). Consequently, they are widely used in stream bio-monitoring, restoration, and predictable to human influences on aquatic systems (Morse et al., 2007; Chen et al., 2013; Adugna and Alemu, 2017).

Many efforts are dedicated to modeling the benthic macroinvertebrate community based on the environment factors. The mathematical modeling with expressions of community dynamics (Gersteva et al., 2004), the hierarchical Bayesian model (Wyatt, 2003), the neural network model (Olden et al., 2006), Decision trees (D'heygere et al., 2003), STELLA model (Li and Yakupitiyage, 2003), RIVPACS-style models (Wright,

DOI: http://dx.doi.org/10.15666/aeer/1601_387406

1995; Hawkins et al., 2000; Davy-Bowker et al., 2008), AUSRIVAS model (Simpson and Norris, 2000) are all applied to the studies. Most of these modeling aim at solving certain function- and process- oriented questions. Some of them are limited to lots of environment variables or available data. Despite all of these studies, the impact of river water quality on the macroinvertebrates community is not clear thoroughly. And the predictive accuracy of the models is inadequate. It hooks the predictive models in using widely. Therefore, we try two models in this study in order to dig the relationship between river water quality and macroinvertebrates deeply, and achieve the satisfactory predicting accuracy (Halim et al., 2017).

Many studies are conducted on the relations of macroinvertebrate communities to environmental factors, using abundance, richness, diversity variables (Clarke et al., 2003; Wyatt, 2003; Bonada et al., 2006; Mereta et al., 2012; Pan et al., 2012; Chen et al., 2013; Jiang et al., 2014; Sarkar et al., 2017). Results of previous studies have indicated that the environmental factors such as conductivity (Mesa, 2010), water temperature (Camur-Elipek et al., 2010), total nitrogen (Couceiro et al., 2007), total phosphorus (Maul et al., 2004), dissolved oxygen (Kaller and Kelso, 2007) and chemical oxygen demand (Song et al., 2009) are the important environmental factors impacted on macroinvertebrate assemblages.

Although, lots of studies on the relations between macroinvertebrate assemblages and environmental factors in aquatic ecosystems are carried out, the scarce of those in city river ecosystems still exists (Hashemi, 2017). Moreover, the prediction of macroinvertebrate assemblages should also be conducted more and deeply, better providing more useful implications for conservation and management of river and stream ecosystems. Thus, it is necessary to carry out quantitative studies on the relations of macroinvertebrate assemblages to hydro-environmental factors in urban rivers.

Therefore, the present study applies two procedures, linear model and k-nearest neighbor method, to predict the biodiversity of macroinvertebrate assembles in a city river, using the water quality indicators. The purposes of this work were: 1) to build macroinvertebrate biodiversity predicted models; 2) to compare the validities of linear model and k-nearest neighbor method.

Materials and methods

Study area

Wenyu River is the only one originating from Beijing urban area. It flows into North Canal through the Beiguan gate dam, located in Tongzhou District (*Figure 1*). There are three tributaries, Dongsha River (flowing through Changping District), Beisha River (flowing through Changping District) and Nansha River (flowing through Haidian District), which conflow at the Shahe Reservoir located in Changping District to form the upstream of Wenyu River with the drainage area of 1099 km^2 (Meng et al., 2010; Xiao et al., 2017; Radan et al., 2017). The segment after Shahe gate dam is described to “Wenyu River”, flowing southeast into Beiguan gate dam, through Changyang District and Shunyi District. It is 47.5 km long, with a drainage area of 2478 km^2 (Meng et al., 2010; Vazdani et al., 2017). The segment from Shahe gate dam to Lutong gate dam is the middle reaches of Wenyu River, with a length of 23 km. Lingou River is the main tributary of the middle reaches. The segment from Lutong gate dam to Beiguan gate dam is called the downstream of Wenyu River, with a length of 24.5 km. Qing River, Ba River and Xiaozhong River contributes the main tributaries of the downstream

(*Figure 1*). The mainstream and the associated riparian areas of Wenyu River are intensively affected by urban developments. Wenyu River is a typical urban river in China, with its segment flowing through many urban lands. The problem of channelization in the river is very severe.

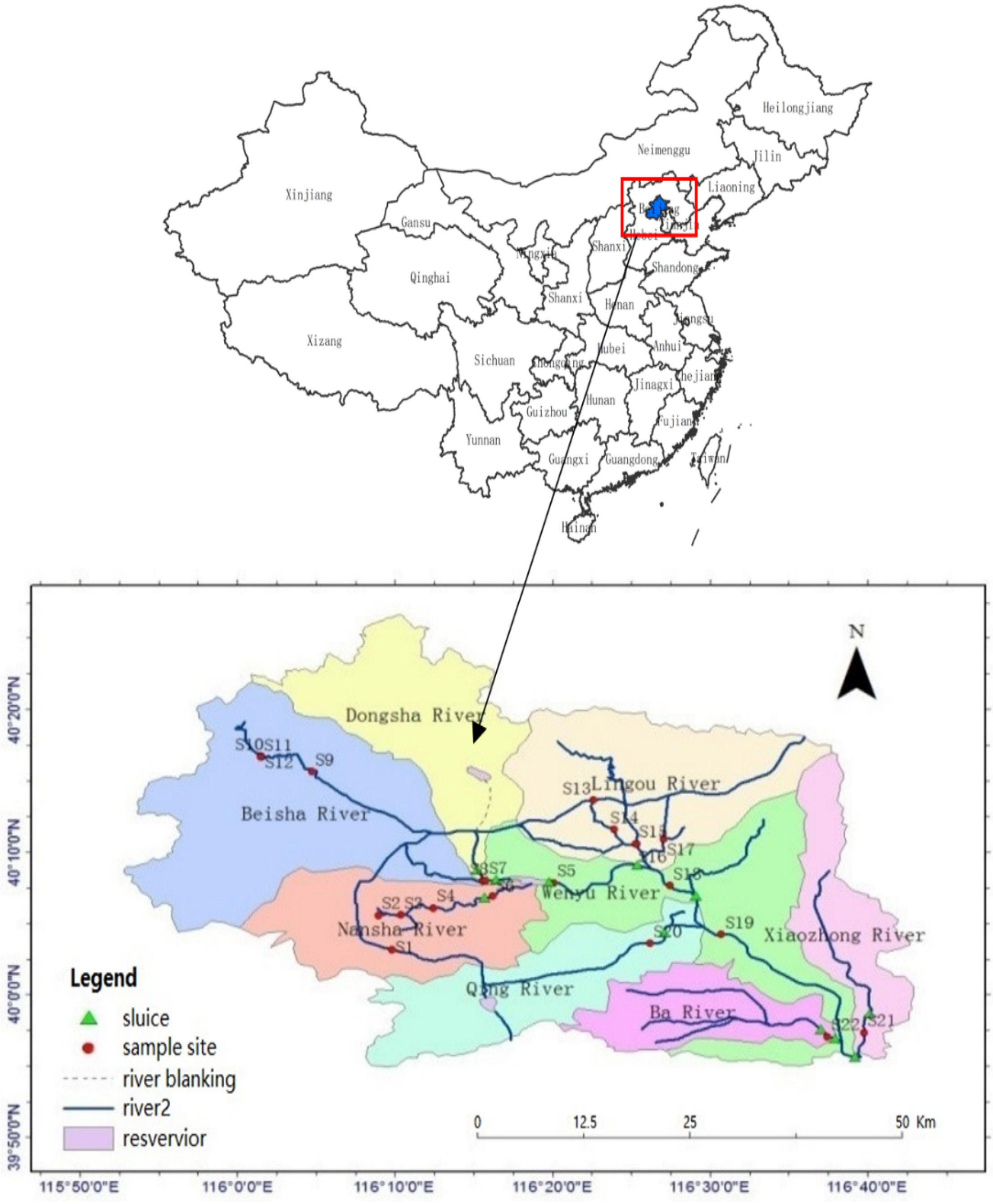

Figure 1. *Locations of the study area and the sample sites in Wenyu River, Beijing*

The drainage of Wenyu River belongs to temperate zone and the climate is continental monsoon climate. The rainfall varies greatly both between years and within one year. The mean annual rainfall is almost 600 mm, with 80% falling in wet season

DOI: http://dx.doi.org/10.15666/aeer/1601_387406

from June to September. The mean annual runoff is almost 350 million m^3 with 60%~70% coming from wastewater (Yang et al., 2014; Xiao et al., 2017).

Sample sites

A total of 22 sampling sites (abbreviated as S1 to S22) are monitored from the upstream tributaries to the downstream (*Figure 1*). Eleven sites are selected from the upstream tributaries (sites S1, S2, S3, S4 and S6 located in Nansha River, sites S7 and S8 located in Dongsha River, sites S9 to S12 located in Beisha River). Seven sites are selected from the middle reaches (sites S13 to S17 located in Lingou River, sites S5 and S18 located in Wenyu upper mainstream). Four sites are selected from the downstream reaches (site S19 located in Wenyu lower mainstream, site S20 located in Qing River, site S21 located in Xiaozhong River, site S22 located in Ba River) (*Figure 1*). The sampling sites selection is restrained by some construction and agricultural activities. For example, S12, located in the upstream reaches, are fenced and no entering because of the villager's fishing or paving cement at the bottom of the river.

Data collection

Water and macroinvertebrates samples are collected in every autumn (October to November) from 2010 to 2012, in each sampling site. Three macroinvertebrates samples are taken by a Peterson grab dredger (1/16 m^2) in each site. The samples are sieved by a 500 μm mesh sieve in situ. The animal individuals are selected from sediment manually on a white porcelain plate and conserved in 75% ethanol for identification. The organisms are identified to species level using a stereoscopic dissection microscope (magnification 10-75×) and counted (Zhou and Chen, 2011; Wang and Wang, 2011; Yang et al., 2014; Yang et al., 2017). Wet weight of macroinvertebrates is determined by an electronic balance after being blotted. The population density (ind/m^2) and biomass density (g/m^2) of each species in each sampling site are calculated respectively.

According to the literature, 12 physical and chemical variables are measured and sampled before macroinvertebrate sampling. Temperature (MYRONL ULTRAMETER Ⅱ6PFC), conductivity (MYRONL ULTRAMETER Ⅱ6PFC), pH (MYRONL ULTRAMETER Ⅱ6PFC), turbidity (HACH 2100N Turbidimeter) and dissolved oxygen (DO) (HACH HQ30d) are measured on site at each sampling site. Water samples for chemical variables analyses are collected by a water sampler and are conserved in 500ml polyethylene bottle at each sampling site. All the water samples are put in an ice chest at 4 °C and are analyzed within 24 h after collection. The total nitrogen (TN) and total phosphorus (TP) are analyzed by UV spectrophotometer. The biochemical oxygen demand (BOD_5) is determined by dilution inoculation method. The chemical oxygen demand (COD_{Mn}) is analyzed by potassium permanganate method. Ammonia nitrogen (NH_3-N), Nitrate nitrogen (NO_3^--N) and Nitrite nitrogen (NO_2^--N) are analyzed by gas phase molecular absorption spectrum method.

Linear model and k-nearest neighbor method

Two methods are utilized to build the relationship model between the biodiversity indices of macroinvertebrates and water quality concentrations, the linear model and the k-nearest neighbor method. Data in 2010 and 2011 are used to build the models and Data in 2012 are used to test the validity of models. The Shannon-Wiener index and

Simpson index are both used in the linear model and the k-nearest neighbor method where the water quality indicators are the explainable variables. The biodiversity indices for the years 2010, 2011 and 2012 are both calculated by R software version 3.1.1, using the collected macroinvertebrate taxa data. We obtain 57 observations after eliminating the default (got samples but had no macroinvertebrates) and empty sample sites (do not obtain samples) (*Table 1*).

Table 1. *Numbers of observations in each sample site*

Sample sites	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**
Numbers of observations	3	3	2	2	3	2	3	3	3	3	3
Sample sites	**S12**	**S13**	**S14**	**S15**	**S16**	**S17**	**S18**	**S19**	**S20**	**S21**	**S22**
Numbers of observations	1	3	3	2	3	3	3	3	2	2	2

Results

Correlations between biodiversity indices and water quality indicators

We firstly compute the correlation matrix of biodiversity indices (Shannon-Wiener index and Simpson index) and the concentrations of 12 water quality indicators (*Table 2*). As far as Shannon-Wiener index is concerned, seven water quality indicators were significantly correlated with it (p-value ≤ 0.05), pH, DO, conductivity, NH_3-N, TP, COD_{Mn} and BOD_5. Whereas for Simpson index, less water quality indicators show significant correlations (p-value ≤ 0.05), only 5 of 12, DO, conductivity, NH_3-N, TP and COD_{Mn} (*Table 2*).

Table 2. *Correlation matrix of biodiversity indices and water quality concentration*

Water quality indicators	**pH**	**DO**	**Temperature**	**Turbidity**	**Conductivity**	**TN**
Shannon-Wiener index	0.276*	0.431*	-0.167	-0.216	-0.498*	-0.187
Simpson index	0.138	0.322*	-0.175	-0.108	-0.415*	-0.156
Water quality indicators	**NH_3-N**	**TP**	**COD_{Mn}**	**BOD_5**	**NO_3^--N**	**NO_2^--N**
Shannon-Wiener index	-0.413*	-0.423*	-0.457*	-0.318*	0.175	-0.056
Simpson index	-0.306*	-0.322*	-0.410*	-0.168	0.171	0.045

*Significant under the significance level of 0.05

Linear model for Shannon-Wiener index

Linear model

Shannon-Wiener index were transformed by $log\left(x+1\right)$, the nature logarithm transformation, since they are nonnegative numbers originally. We then used the R function lm() to fit the model (*Eq. 1*), which is:

$$\begin{aligned}\log(\mathrm{y}_i+1) = {} & \beta_0 + \mathrm{p}H_i \times \beta_1 + DO_i \times \beta_2 + Temperature_i \times \beta_3 + Turbidity_i \times \beta_4 + \\ & Conductivity_i \times \beta_5 + TN_i \times \beta_6 + NH_3 - N_i \times \beta_7 + TP_i \times \beta_8 + \\ & CODmn_i \times \beta_9 + BOD_{5i} \times \beta_{10} + NO_3^- - N_i \times \beta_{11} + NO_2^- - N_i \times \beta_{12} + \varepsilon_i\end{aligned} \tag{Eq. 1}$$

DOI: http://dx.doi.org/10.15666/aeer/1601_387406

where y_i is the Shannon-Wiener index of observation, $i = 1, 2, \cdots, 38$, $\varepsilon_i i.i.d \quad N(0, \sigma^2)$ and σ^2 is unknown.

There are 12 variables and 38 observations in the model. Considering that not all the predictor variables are correlated to the response, we select the variables by AIC in a stepwise algorithm which is implemented by R function step, then we had the linear model (*Eq. 2*).

$$\begin{aligned}\log(y_i + 1) = 1.506 - 0.014 \times Temperature_i - 0.454 \times Conductivity_i - \\ 0.025 \times COD_{Mn_i} - 0.027 \times NO_3^- - N_i + 0.185 \times NO_2^- - N_i + \varepsilon_i\end{aligned} \quad \text{(Eq. 2)}$$

Figure 2 shows the observed values and fitted values:

$$\begin{aligned}(\log(\mathrm{y}_i + 1) = 1.506 - 0.014 \times Temperature_i - 0.454 \times Conductivity_i - \\ 0.025 \times CODmn_i - 0.027 \times NO_3^- - N_i + 0.185 \times NO_2^- - N_i)\end{aligned}$$

for Shannon-Wiener $\log(y_i + 1)$ of 38 observations. The variance estimation of the residual $\sigma^2 = 0.142$. The regression model's adjusted $R^2 = 0.588$. The more the adjusted R^2 is, the better the fitness of linear model is. The fitted results show the moderate correlations.

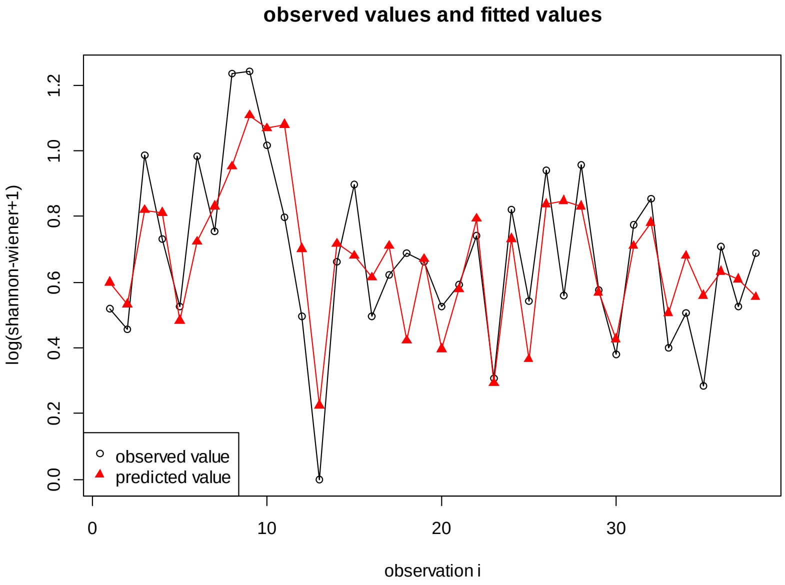

Figure 2. *Plot of observed values and fitted values of Shannon-Wiener index by linear model*

Validation of the prediction model

Applying the linear model (*Eq. 2*), Shannon-Wiener index of 22 sample sites are predicted by 12 concentrations of water quality indicators in Wenyu River monitored in 2012. They are compared to those computed by macroinvertebrate assembles samples collected at the same period (*Figure 3*). Since the Shannon-Wiener index takes non-negative numbers, the fitted values and 0 are assigned by max (*Eq. 3*). The 99%

confidence intervals of the predicted Shannon-Wiener index are presented in *Figure 3*. It shows that there are 68.42% observations within the 99% confidence interval. For a given new sample, the predicted value is:

$$\begin{aligned}&y_{new,i} = \max(\exp\{1.506 - 0.014\times Temperature_{new,i} - 0.454\times Conductivity_{new,i} - 0.025\times \\ &CODmn_{new,i} - 0.027\times NO_3^- - N_{new,i} + 0.185\times NO_2^- - N_{new,i}\} - 1, 0) \\ &\hat{y}_{new} = \max\end{aligned} \quad \text{(Eq. 3)}$$

where, $\hat{y}_{new}$ is the predicted Shannon-Wiener index of 22 sample sites in Wenyu River in 2012, $i = 1, 2, \cdots, 22$.

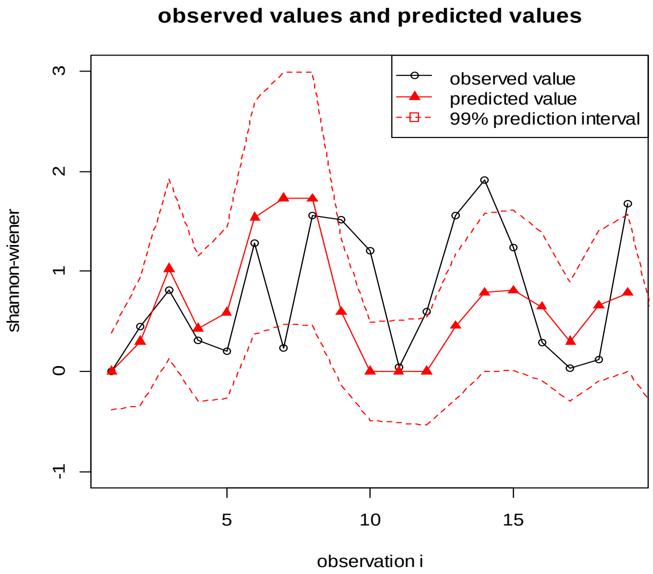

Figure 3. *The 99% prediction interval for Shannon-Wiener index using Linear model*

K-nearest neighbor method for Shannon-Wiener index

K-nearest neighbor method

The k-nearest neighbor method uses the points being close to the point of interest to do the training and predicting, where the Mahalanobis distance (*Eq. 4*) is used to evaluate the quantity of the closeness.

$$D(z_1, z_2) = \sqrt{(z_1 - z_2)^T S^{-1}(z_1 - z_2)} \quad \text{(Eq. 4)}$$

where, z_1 and z_2 are p-dimensional column vectors, and S is the covariance matrix of z_1 and z_2. Here p=12.

The mean of Shannon-Wiener index of the observations is used as the predictive result. Denoted the set of the points that are closed to the points of interest by $\Omega(x_0)$. Let x_i denotes the observations of the above 12 features. Then the predicted value of the Shannon-Wiener index is calculated by (*Eq. 5*):

DOI: http://dx.doi.org/10.15666/aeer/1601_387406

$$y(x_0) = \frac{1}{\#\Omega(x_0)} \sum_{i \in \Omega(x_0)} y(x_i) \qquad \text{(Eq. 5)}$$

where, $\#\Omega(x_0)$ denotes the number of the points in the set $\Omega(x_0)$. For a given is δ with $\delta > 0$, $\Omega(x_0)$ is calculated by (*Eq. 6*):

$$\Omega(x_0) = \{i : D(x_i, x_0) \le \delta\} \qquad \text{(Eq. 6)}$$

We use the cross validation method to find the optimal *k*, the number of points in $\Omega(x_0)$. The whole 38 observations are randomly partitioned into 5 subsamples, and the l^{th} subsample has n_l observations. A subsample is retained as the testing data for testing the model, and the remaining 4 subsamples are used as training data for fitting the model for each time. Then we obtain the predicted value for each observation, and use the mean squared prediction error to determine the optimal *k* that makes the mean squared prediction error being the smallest.

We obtain the predicted values by use of different *k* (1,2,…,20). Furthermore, we estimate the mean prediction error by *Equation 7* and get the line graph (*Figure 4*).

$$MSPE = \frac{1}{38} \sum_{l=1}^{5} \sum_{i=1}^{n_l} \left(y_{li} - y_{li}\right)^2 \qquad \text{(Eq. 7)}$$

where, y_{li} and y_{li} denotes the original values and the predicted values at the *i*th site in l^{th} subsample.

According to *Figure 4,* the mean prediction error is least when *k* is 7.

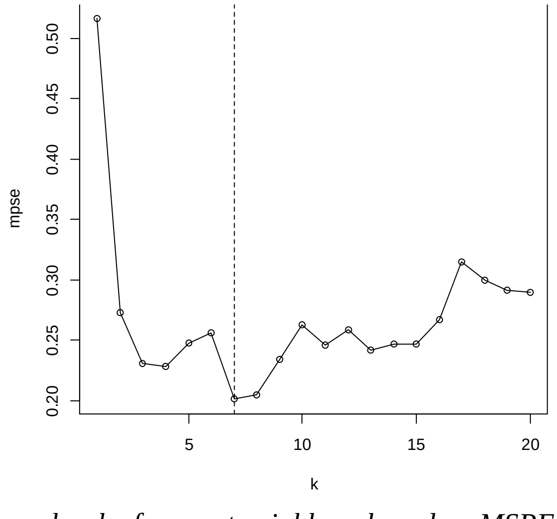

Figure 4. *The optimal number k of nearest neighbors based on MSPE using cross validation*

Prediction error of the method

Therefore, we set $k = 7$ to estimate the Shannon-Wiener index of the 22 sample sites of 2012, using the data of years 2010 and 2011. The 99% prediction interval for

Shannon-Wiener index using k-nearest neighbors is given in *Figure 5*. It showed that there are 42.11% observations within the 99% confidence interval. The prediction validity is obviously not good.

Linear model for Simpson index

Linear model

We use the same analysis for Simpson index. The Simpson index and Shannon-Wiener index are significantly positive correlated (Corr(simpson, Shannon) = 0.886).

Using the data of 2010 and 2011, the fitted model is estimated by (*Eq. 8*):

$$\begin{aligned}\log(y_i+1)=\beta_0+pH_i\times\beta_1+DO_i\times\beta_2+Temperature_i\times\beta_3+Turbidity_i\times\beta_4+\\ Conductivity_i\times\beta_5+TN_i\times\beta_6+NH_3-N_i\times\beta_7+TP_i\times\beta_8+COD_{Mn_i}\times\beta_9+\\ BOD_{5i}\times\beta_{10}+NO_3^{-}-N_i\times\beta_{11}+NO_2^{-}-N_i\times\beta_{12}+\varepsilon_i,\end{aligned} \qquad \text{(Eq. 8)}$$

where, y_i is the Simpson index of observation of the number i. $\varepsilon_i i.i.d \quad N\left(0,\sigma^2\right)$, σ^2 is unknown.

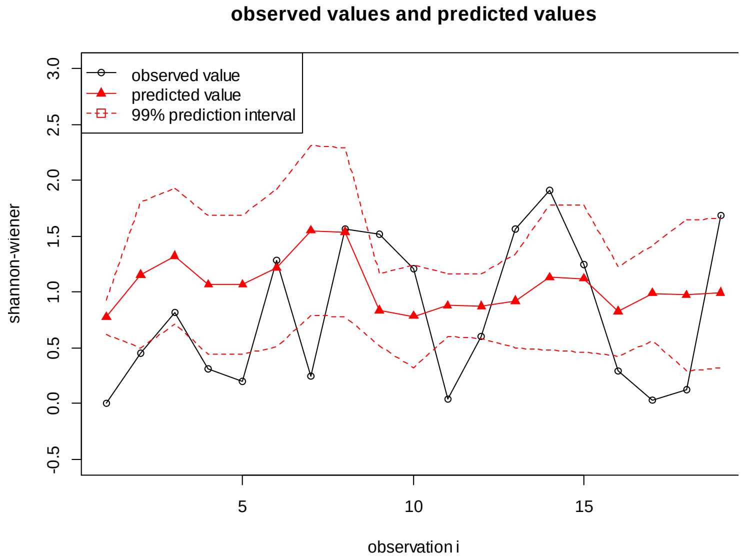

Figure 5. *The 99% prediction interval for Shannon-Wiener index using k-nearest neighbors*

We also select the variables by AIC in a stepwise algorithm which is implemented by R function, then we have the linear model (*Eq. 9*), considering that not all the predictor variables are correlated to the response.

$$\begin{aligned}\log(y_i+1)=0.819-0.344\times Conductivity_i-0.009\times COD_{Mn_i}-0.020\times NO_3^{-}-\\ N_i+0.160\times NO_2^{-}-N_i\end{aligned} \qquad \text{(Eq. 9)}$$

where, y_i is the Simpson index of observation, $i=1,2,\cdots,38$. $\varepsilon_i i.i.d \quad N\left(0,\sigma^2\right)$, σ^2 is unknown.

We get the comparison plot of observed values and fitted values:

$$(\log(\ y_i + 1\ ) = 0.819 - 0.344 \times Conductivity_i - 0.009 \times COD_{Mn_i} - 0.020 \times NO_3^- - N_i + 0.160 \times NO_2^- - N_i)$$

for Simpson $\log(y_i + 1)$ of 38 observations (*Figure 6*). The variance estimation of the residual $\sigma^2 = 0.016$. The regression model's adjusted $R^2 = 0.394$.

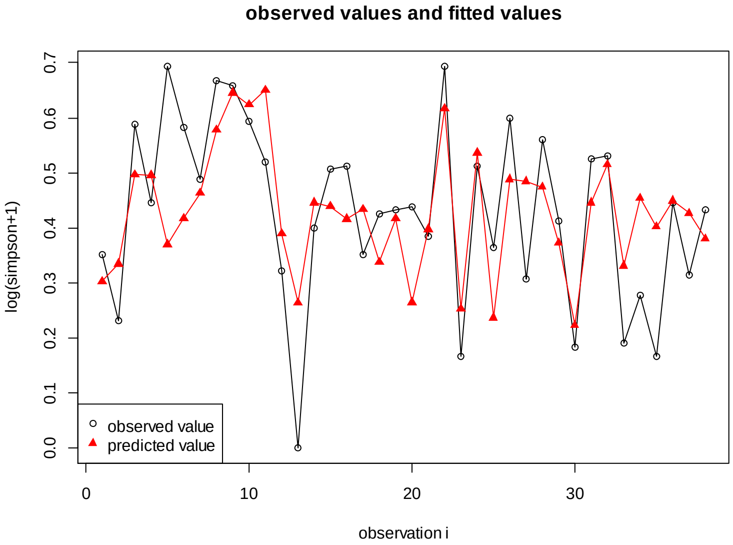

Figure 6. *Plot of observed values and fitted values of Simpson index by Linear model*

Validation of the prediction model

Applying the linear model (*Eq. 9*), Simpson index of 22 sample sites are predicted by 12 concentrations of water quality indicators in Wenyu River monitored in 2012. They are compared to the actual measured Simpson index at the same period (*Figure 7*).

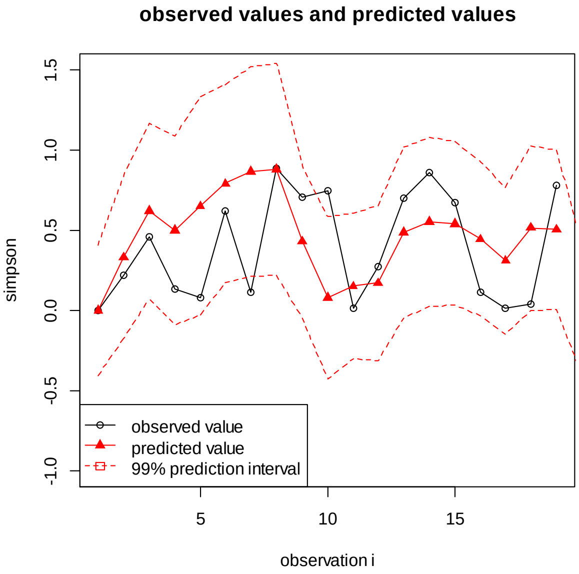

Figure 7. *The 99% prediction interval for Simpson index using Linear model*

DOI: http://dx.doi.org/10.15666/aeer/1601_387406

Since the Simpson index is also non-negative numbers, the fitted values and 0 are assigned by max again (*Eq. 10*). A 99% confidence interval of the predicted Simpson index is presented in *Figure 8*. There are 89.47% observations within the 99% confidence interval.

$$y_{new,i} = \min\{\max(\exp\{0.819 - 0.344 \times Conductivity_i - 0.009 \times COD_{Mn_i} - 0.020 \times NO_3^- \\ -N_i + 0.160 \times NO_2^- - N_i\} - 1, 0), 1\} \quad \text{(Eq. 10)}$$

where, $y_{new,i}$ is the predicted Simpson index of 22 sample sites in Wenyu River in 2012 $i = 1, \cdots, 22$.

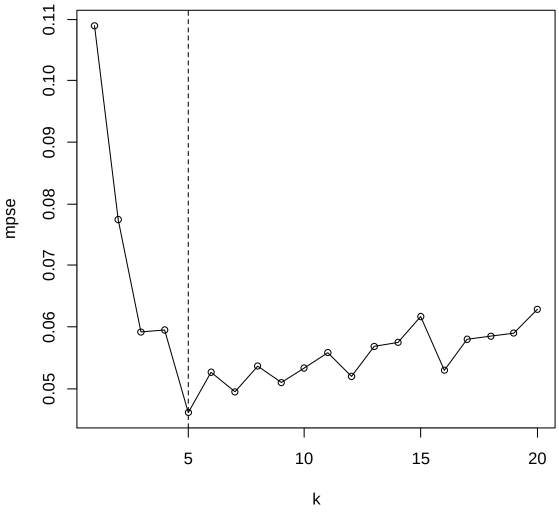

Figure 8. *The optimal number k of nearest neighbors based on PSME using cross validation*

K-nearest neighbor method for Simpson index

Similar with Shannon-Wiener index, we use the cross validation method to find the optimal *k* for Simpson index. The whole 38 observations are randomly partitioned into 5 subsamples, and the l^{th} subsample has n_l observations. We obtain the predicted value for each observation, and use the mean squared prediction error to determine the optimal *k* for which the mean squared prediction error is the smallest.

We obtain the predicted values by using different *k*(1,2,…,20). Furthermore, we estimate the mean prediction error by *Equation 7* and get the line graph (*Figure 8*). According to *Figure 8,* the mean prediction error is found to be the smallest when *k* is 5.

Therefore, we set $k = 5$ when we estimate the Simpson index of the 19 sample sites of 2012, using the data of two former years. The observed values and fitted values are compared to test the validity of 5-nearest neighbor method. A 99% confidence interval of the predicted Simpson index by 5-nearest neighbors is presented in *Figure 9*. There are 21.05% observations within the 99% confidence interval.

Comparisons of different simulated methods and biodiversity index

We put the predicted values by linear model and k-nearest neighbor method together in one plot, in order to compare the validities of two methods (*Figure 10*). According to *Figure 10,* the result of the linear model is better than the other method for Shannon-

Wiener index. The test results of 99% prediction intervals also show the same conclusion. There are 68.42% observations within the prediction intervals by linear model, while only 42.11% are within the prediction intervals by k-nearest neighbor method.

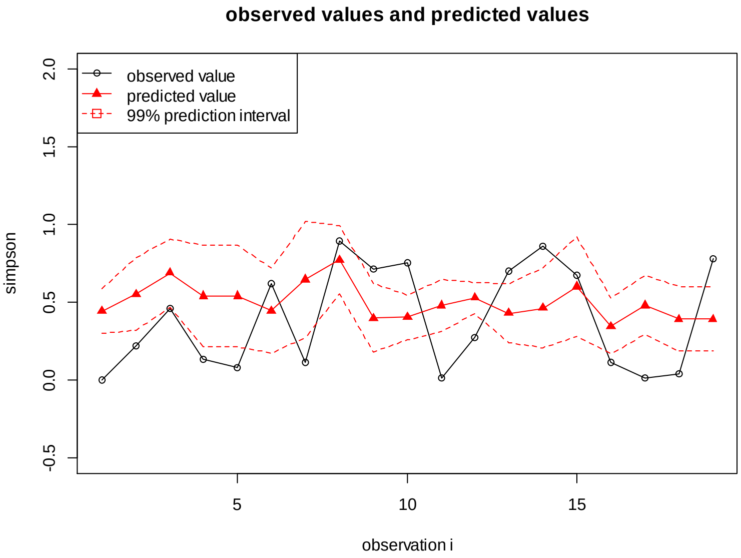

Figure 9. *The 99% prediction interval for Simpson index using k-nearest neighbors*

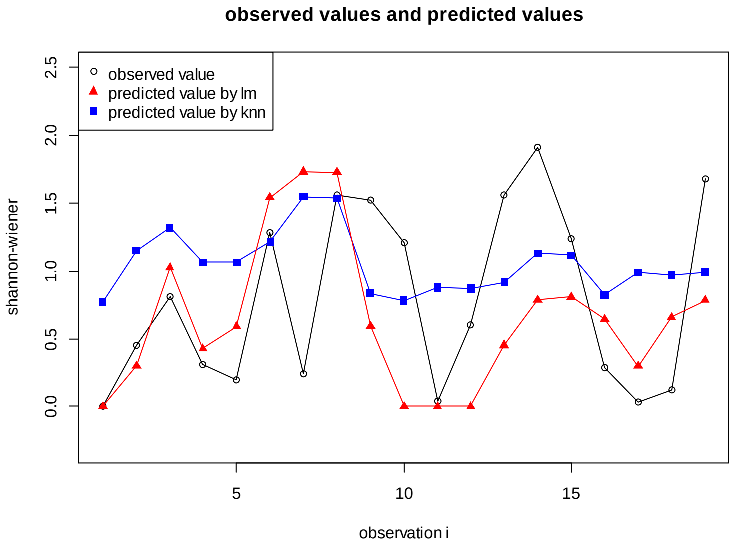

Figure 10. *Plot of Shannon-Wiener index predicted values by linear model and k-nearest neighbor method*

Similar conclusion is found for Simpson index prediction. The linear model (89.47% within the prediction intervals) is more suitable for predicting Simpson index than k-nearest neighbor method (21.05% within the prediction intervals) in Wenyu River (*Figure 11*).

As for the different biodiversity indices, Simpson index show more appropriate than Shannon-Wiener index for predicting macroinvertebrate assembles using water quality indicators in Wenyu River, a typical city river (*Figure 10, Figure 11*). There are 89.47%

observations within the 99% confidence interval for Simpson index, whereas 68.42% for Shannon-Wiener index.

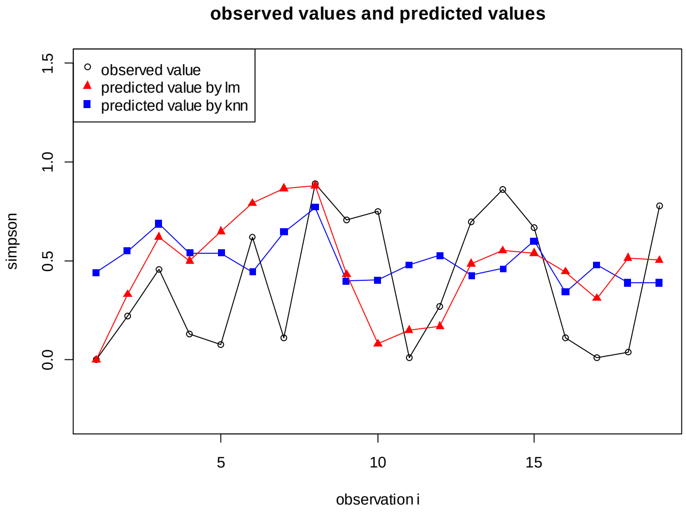

Figure 11. *Plot of Simpson index predicted values by linear model and k-nearest neighbor method*

Discussion

Biomass of macroinvertebrates

We chose the biodiversity index as the variable of macroinvertebrates. However, the abundance and biomass are often applied to researches conducted on the relationship between water quality and benthic macroinvertebrates in river systems. We also try to make macroinvertebrates prediction model using abundance and biomass variables. It is a pity that these two common variables show almost the same depressing predicted results. Abundance and biomass, thereby, are abandoned in this study. Considering the article’s length, we only take biomass as an example to explain the depressing result.

y_{it} is the total benthic macroinvertebrate biomass of the number t sampling of the i^{th} sample site in Wenyu River. $pH_{it},\cdots,NO_2^- - N_{it}$ is the concentration of 12 water quality indicators of the number t sampling of the i^{th} sample site, $i=1,2,\cdots,22, t=1,\cdots,n_i$.

P value is less than 0.01 when y_{it} is in the normality of test. Therefore, y_{it} is made a transformation by box cox, $\lambda = 0.107$, which is close to 0, similar to the transformation by $\log(y)$. We get the histogram of $\log(y_{it})$ (*Figure 12*). We find $\log(y_{it})$ to be following the normal distribution approximately.

The relation model of the total biomass of macroinvertebrates and water quality indicators is found by (*Eq. 11*):

$$\begin{aligned}\frac{\log(y_{it})-\overline{y}}{S_y} &= \beta_0 + pH_{it}\times\beta_1 + DO_{it}\times\beta_2 + Temperature_{it}\times\beta_3 + Turbidity_{it}\times\beta_4 + \\ & Conductivity_{it}\times\beta_5 + TN_{it}\times\beta_6 + NH_3 - N_{it}\times\beta_7 + TP_{it}\times\beta_8 + \\ & CODMn_{it}\times\beta_9 + BOD_{5it}\times\beta_{10} + NO_3^- - N_{it}\times\beta_{11} + NO_2^- - N_{it}\times\beta_{12} + \\ & pH_{it}^{\ 2}\times\beta_{13} + DO_{it}^{\ 2}\times\beta_{14} + \cdots + NO_3^- - N_{it}\times NO_2^- - N_{it}\times\beta_{90} + u_i + \varepsilon_{it}\end{aligned} \qquad \text{(Eq. 11)}$$

DOI: http://dx.doi.org/10.15666/aeer/1601_387406

where, $\bar{y}=\frac{1}{260}\sum_{i=1}^{22}\sum_{t=1}^{n_i}\log(y_{it})$, $S_y{}^2=\frac{1}{259}\sum_{i=1}^{22}\sum_{t=1}^{n_i}(\log(y_{it})-\bar{y})^2$. y_{it} is the total benthic macroinvertebrate biomass of the number t sampling of the i^{th} sample site in Wenyu River. $pH_{it},\cdots,NO_2{}^{-}-N_{it}$ - the concentration of 12 water quality indicators of the number t sampling of the i^{th} sample site, $i=1,2,\cdots,22,t=1,\cdots n_i$.

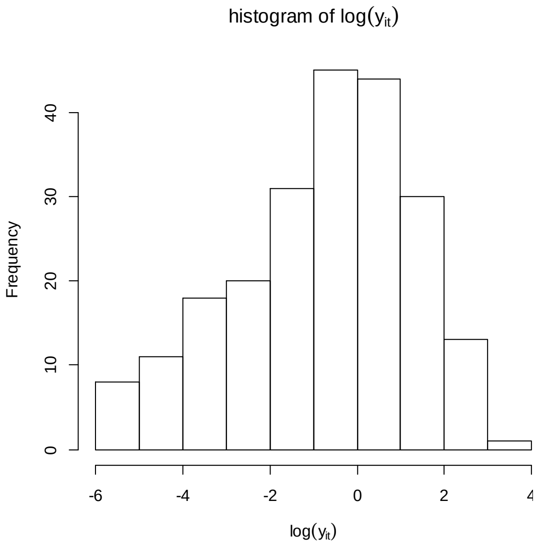

Figure 12. *Histogram of the normality of test of* $\log(y_{it})$. y_{it} *- Total benthic macroinvertebrate biomass of the number t sampling of the* i^{th} *sample site in Wenyu River*

We set values for λ from 0 to 26 (when $\lambda=26$, all the variables are not considered in the model), and maximized equation (*Eq. 12*). The Akaike information criterion (AIC) is the minimum when $\lambda=10$ (*Figure 13*). We get the non-negative variable β_0= -0.0538 and the final line mixed model (*Eq. 13*).

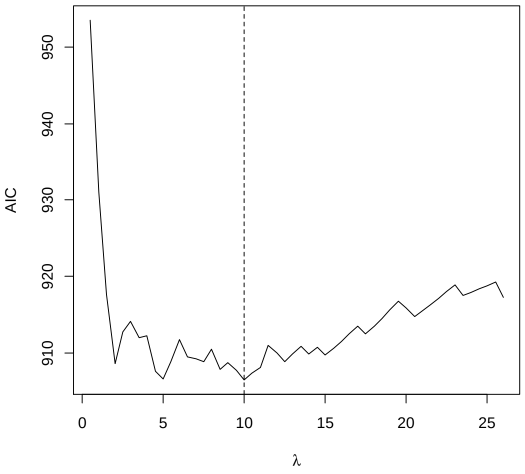

Figure 13. *The line graph of the model AIC and* λ

DOI: http://dx.doi.org/10.15666/aeer/1601_387406

$$Q(\beta,\sigma_u,\sigma_\varepsilon)=\frac{1}{2}\log(\det(V))+\frac{1}{2}(y-X\beta)^T V^{-1}(y-X\beta)+\lambda\sum_{k=1}^{90}|\beta_k| \qquad \text{(Eq. 12)}$$

$$\begin{aligned} y_{it} = & -0.14446+0.0793PH_{it}-0.0006NH4_{it}{}^2-0.0018PH_{it}\times NH4_{it}+ \\ & 0.003DO_{it}\times Temp_{it}+0.0199DO_{it}\times Cond_{it}-0.0002Tureb_{it}\times COD_{Mn_{it}}- \\ & 0.0002NH4_{it}\times COD_{Mn_{it}}+0.0155NO3_{it}\times NO2_{it}+u_{it}+\varepsilon_{it} \end{aligned} \qquad \text{(Eq. 13)}$$

where, $u_i i.i.d \quad N(0,0.5486)$, $\varepsilon_{it} i.i.d \quad (0,2.8843)$.

The Tukey-Anscombe residual plot (*Figure14*) and QQ plot of the residuals (*Figure 15*) show that the residuals of the linear mixed model accorded with normal distribution.

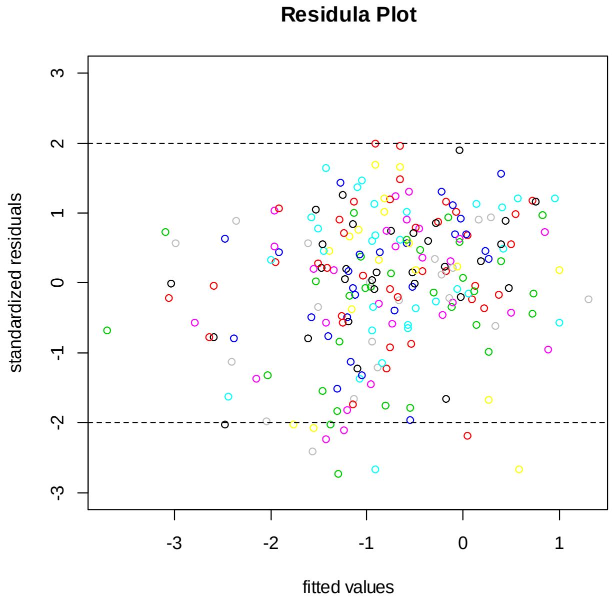

Figure 14. *Tukey-Anscombe residual plot of fitted values*

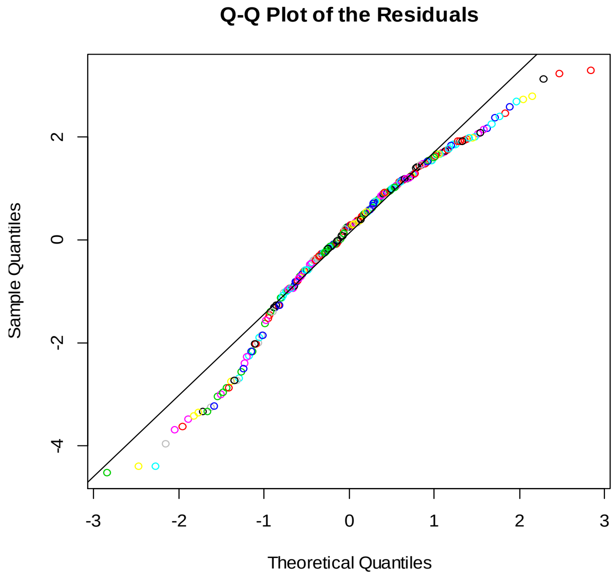

Figure 15. *QQ plot of the residuals*

DOI: http://dx.doi.org/10.15666/aeer/1601_387406

Normal distribution is also presented in the QQ plot of the standardized random effects (*Figure 16*). However, the prediction displayed an inaccurate result, contrast to the observed biomass values, by the line mixed model (*Figure 17*). Therefore, the biomass index is given up. The reason of this is yet not clear. It is perhaps concerned with the unpleasant water quality all through the river.

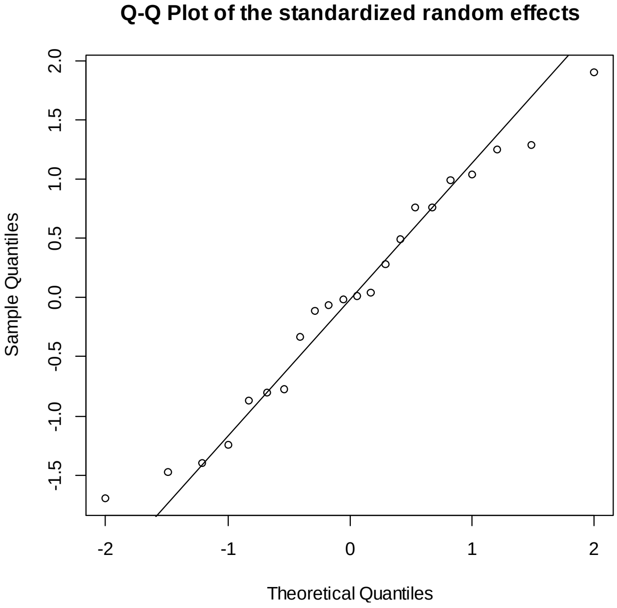

Figure 16. *QQ plot of the standardized random effects*

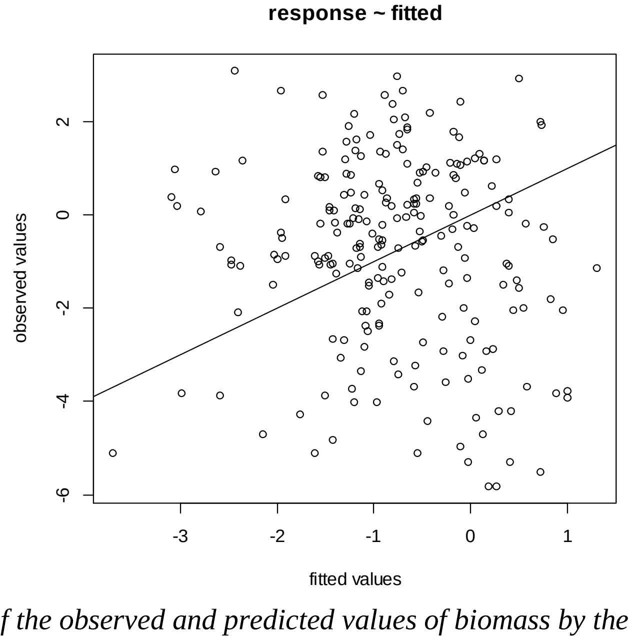

Figure 17. *Plot of the observed and predicted values of biomass by the line mixed model*

Simpson index

Simpson index shows a better predicted result than Shannon-Wiener index in this study. We know that Simpson index is more sensitive to the evenness index of a community while Shannon-Wiener index is more sensitive to the abundance index (Ma et al., 1995). The abundance of benthic macroinvertebrates in each sample site varies from 1 to 14, which 1 and 2 species most frequently appeared in sample sites. The abundance index changing distinctively accounted for Simpson index's better than Shannon-Wiener index.

Another reason about this could perhaps found in the research of Magurran (1988). He claims that Simpson index is more sensitive to the dominant species the Shannon-Wiener index. It seems the case in our study. For example, there are two species of macroinvertebrates in the sample site S19. One is *Limnodrilusclaparedianus*, the other is *Branchiurasowerbyi*. The individuals of the former are 3216 whereas the latter is 1. The similar status appears in most of the sample sites. The dominant species have apparent superiority of the amount and thus could affect the result of biodiversity prediction.

Limitations

We conduct a study to predict biodiversity of macroinvertebrates in a city river using two biodiversity index and 12 water quality indicators. Unfortunately, there are only 57 observations used in total, in which 38 are used for model training and 19 for validation. The poor data quality maybe affects the accurate conclusion about the prevalence of the linear model over the KNN. We should accumulate more and more observations during the next years for the supplement comparison study of these two models.

We use 12 water quality indicators to get the correlations with the biodiversity index of benthic macroinvertebrates. However, riverbed substrate and flow velocity have also important effect on macroinvertebrates (Damanik-Ambarita et al., 2016; Berger et al., 2017). Since the flow velocity of the observations in Wenyu river has little difference from each other, there is no significant correlation between flow velocity and biodiversity index in Wenyu river. Riverbed substrate types should be discussed in the future studies.

Conclusion

In this study we build two models, a linear model and k-nearest neighbor method, to predict biodiversity of macroinvertebrates in Wenyu river from the measured data of water and macroinvertebrates. Furthermore, the predicting ability of these two models are compared. We find the linear model is better for predicting macroinvertebrates diversity using water quality indicators than k-nearest neighbor method. For biodiversity indicators, Simpson index appears more robust and accurate than Shannon index for predicting benthic macroinvertebrates in a city river. The developed predictive model indicates a useful tool for assessing river health, especially city river health, since there were 89.47% observations within the 99% confidence intervals. The results of this paper could do some help to river health assessment and management.

Acknowledgement. We thank the many people involved in collecting and processing the physicochemical and macroinvertebrate samples. Particularly, we thank Wang Jun-cai for identifying benthic macroinvertebrats and Huang Yue-ying for her valuable help. Financial support for this project was partly provided by Chinese National Administration of Surveying, Mapping and Geoinformation.

REFERENCES

[1] Adugna, O., Alemu, D. (2017): Evaluation of brush wood with stone check dam on gully rehabilitation. – Journal CleanWAS 1(2): 10-13.

[2] Bae, Y. J., Kil, H. K., Bae, K. S. (2005): Benthic macroinvertebrates for uses in stream biomonitoring and restoration. – KSCE Journal of Civil Engineering 9: 55-63.

[3] Barbour, M. T., Gerritsen, J., Snyder, B. D, et al. (eds.) (1999): Rapid Bioassessment Protocols for Use in Stream and Wadeable Rivers: Periphyton, Benthic Macroinvertebrates and Fish. Second Edition. – EPA 841-B-99-002. U.S. Environmental Protection Agency; Office of Water; Washington, D. C.

[4] Berger, E., Haase, P., Kuemmerlen, M., Leps, M., Schafer, R. B., Sundermann, A. (2017): Water quality variables and pollution sources shaping stream macroinvertebrate communities. – Science of The Total Environment 587-588: 1-10.

[5] Bonada, N., Prat, N., Resh, V. H., Statzner, B. (2006): Developments in aquatic insect biomonitoring: a comparative analysis of recent approaches. – Annual Review of Entomology 51: 495-523.

[6] Camur-Elipek, B., Arslan, N., Kirgiz, T., Oterler, B., Guher, H., Ozkan, N. (2010): Analysis of benthic macroinvertebrates in relation to environmental variables of Lake Gala, a national Park of Turkey. – Turkish Journal of Fisheries and Aquatic Sciences 10: 235-243.

[7] Chen, Q. W., Yang, Q. R., Li, R. N., Ma, J. F. (2013): Spring micro-distribution of macroinvertebrate in relation to hydro-environmental factors in the Lijiang River, China. – Journal of Hydro-Environment Research 7: 103-112.

[8] Clarke, R. T., Wright, J. F., Furse, M. T. (2003): RIVPACS models for predicting the expected macroinvertebrate fauna and assessing the ecological quality of rivers. – Eco Model 160: 219-233.

[9] Couceiro, S. R. M., Hamada, N., Luz, S. L. B., Forsberg, B. R., Pimentel, T. P. (2007): Deforestation and sewage effects on aquatic macroinvertebrates in urban streams in Manaus, Amazonas, Brazil. – Hydrobiologia 575: 271-284.

[10] D'heygere, T., Goethals, P. L. M., De Pauw, N. (2003): Use of genetic algorithms to select input variables in decision tree models for the prediction of benthic macroinvertebrates. – Eco Model 160: 291-300.

[11] Damanik-Ambarita, M. N., Lock, K., Boets, P., Everaert, G., Nguyen, T. H. T., Forio, M. A. E., Musonge, P. L. S., Suhareva, N., Bennetsen, E., Landuyt, D., Dominguez-Granda, L., Goethals, P. L. M. (2016): Ecological water quality analysis of the Guayas river basin (Ecuador) based on macroinvertebrates indices. – Limnologica - Ecology and Management of Inland Waters 57: 27-59.

[12] Davies, P. J., Wright, I., A., Findlay, S. J., Jonasson, O. J., Burgin, S. (2010): Impact of urban development on aquatic macroinvertebrates in south eastern Australia: degradation of in-stream habitats and comparison with non-urban streams. – Aquat Ecol 44: 685-700. DOI 10.1007/s10452-009-9307-y.

[13] Davy-Bowker, J., Clarke, R., Corbin, T., Vincent, H., Pretty, J., Hawczak, A., Blackburn, J., Murphy, J., Jones, I. (2008): River Invertebrate Classification Tool. – SNIFFER, Edinburgh.

[14] Halim, H., Abdullah, R., Nor, M. J. M., Aziz, H. A., Rahman, N. A. (2017): Comparison between measured traffic noise in Klang Valley, Malaysia and existing prediction models. – Engineering Heritage Journal 1(2): 10-14.

[15] Hashemi, N. (2017): Recognizing the potential of sustainable use of pasture resources in south khorasan province with approch of carrying capacity. Environment Ecosystem Science, 1(2): 09-12.

[16] Hawkins, C. P., Norris, R. H., Hogue, J. N., Feminella, J. W. (2000): Development and evaluation of predictive models for measuring the biological integrity of streams. – Ecol Appl 10: 1456-1477.

[17] Heino, J., Muotka, T., Mykra, H., Paavola, R., Hamalainen, H., Koskenniemi, E. (2003): Defining macro invertebrate assemblage types of headwater streams: implications for bioassessment and conservation. – Ecological Applications 13: 842-852.

[18] Hejazi, S.M., Lotfi, F., Fashandi, H., Alirezazadeh, A. (2017): Serishm: an eco-friendly and biodegradable flame retardant for fabrics. Environment Ecosystem Science, 1(2): 05-08.

[19] Jiang, X. M., Song, Z. Y., Xiong, J., Xie, Z. C. (2014): Can excluding non-insect taxa from stream macroinvertebrate surveys enhance the sensitivity of taxonomic distinctness indices to human disturbance. – Ecological Indicators 41: 175-182.

[20] Li, L., Yakupitiyage, A. (2003): A model for food nutrient dynamics of semi-intensive pond fish culture. – Aquacult Eng 27 (1): 9-38.

[21] Ma, K., Huang, J., Yu, S., et al. (1995): Research on the diversity of plants communities in Dongling Mountain area of Beijing. – Acta Ecologica Sinica 16(3): 225-234 (In Chinese).

[22] Magurran, A. E. (1988): Ecological Diversity and Measurement. – Princeton University Press, New Jersey.

[23] Maul, J. D., Farris, J. L., Milam, C. D., Cooper, C. M., Testa, S., Feldman, D. L. (2004): The influence of stream habitat and water quality on macroinvertebrate communities in degraded streams of northwest Mississippi. – Hydrobiologia 518: 79-94.

[24] Meng, Q. Y., Li, Q. J., Wang, P. J., Liu, C. (2010): Research on Control Techniques for Pollution Source in Watershed of Wenyu River, pp. 4-12. – China Water Power Press, Beijing (in Chinese).

[25] Mereta, S. T., Boets, P., Bayih, A. A., Malu, A., Ephrem, Z., Sisay, A., Endale, H., Yitbarek, M., Jemal, A., De Mester, L., Goethals, P. L. M. (2012): Analysis of environmental factors determining the abundance and diversity of macroinvertebrate taxa in natural wetlands of Southwest Ethiopia. – Eco Info 7: 52-61.

[26] Mesa, L. M. (2010): Hydraulic parameters and longitudinal distribution of macroinvertebrates in a subtropical Andean basin. – Interciencia 35: 759-764.

[27] Olden, J. D., Joy, M. K., Death, R. G. (2006): Rediscovering the species in community wide predictive modeling. – Ecol Appl 16: 1449-1460.

[28] Pan, B. Z., Wang, Z. Y., Xu, M. Z., Xing, L. H. (2012): Relation between stream habitat conditions and macroinvertebrate assemblages in three Chinese rivers. – Quaternary International 282: 178-183.

[29] Radan, A., Latifi, M., Moshtaghie, M., Ahmadi, M., Omidi, M. (2017): Determining the Sensitive Conservative Site in Kolah Ghazi National Park, Iran, In Order to Management Wildlife by Using GIS Software. Environment Ecosystem Science, 1(2): 13-15.

[30] Sarkar, M. I., Islam, M. N., Jahan, A., Islam, A., Biswas, J. C. (2017): Rice straw as a source of potassium for wetland rice cultivation. – Geology, Ecology, and Landscapes 1(3): 184-189.

[31] Simpson, J. C., Norris, R. H. (2000): Biological Assessment of River Quality: Development of AURIVAS Models and Outputs. – In: Wright, J. E., Sutcliffe, D. W., Furse, M. T. (eds.) Assessing the Biological Quality of Fresh Waters: Rivpacs and Other Techniques. Freshwater Biological Association, Ambleside, UK.

[32] Song, M. Y., Leprieur, F., Thomas, A., Lek-Ang, S., Chon, T. S., Lek, S. (2009): Impact of agricultural land use on aquatic insect assemblages in the Garonne river catchment (SW France). – Aquatic Ecology 4: 999-1009.

[33] Vazdani, S., Sabzghabaei, G., Dashti, S., Cheraghi, M., Alizadeh, R., Hemmati, A. (2017): Fmea Techniques Used in Environmental Risk Assessment. Environment Ecosystem Science, 1(2): 16-18.

[34] Wang, J., Wang, X. (2011): The Chironomid Larvae in the North of China. – Chinese Yanshi Press, Beijing.

[35] Wright, J. F. (1995): Development and use of a system for predicting the macroinvertebrate fauna in flowing waters. – Aust J Ecol 20: 181-197.

[36] Wyatt, R. J. (2003): Mapping the abundance of riverine fish populations: integrating hierarchical Bayesian models with a geographic information system (GIS). – Can J Fish Aquat Sci 60: 997-1006.

[37] Xiao, H., Wang, M., Sheng, S. (2017): Spatial evolution of URNCL and response of ecological security: a case study on Foshan City. – Geology, Ecology, and Landscapes 1(3): 190-196.

[38] Xiao, H., Wang, M., Sheng, S. (2017): Spatial evolution of URNCL and response of ecological security: a case study on Foshan City. Geology, Ecology, and Landscapes, 1(3): 190-196.

[39] Yang, L., Wang, J. C., Bai, X. (2014): Temporal and Spatial Distributions of Macroinvertebrates and Their Atlas in Wenyu River, pp. 8-33. – Science Press, Beijing (in Chinese).

[40] Yang, S., Li, J., Song, Y. (2017): Application of surfactant Tween 80 to enhance Fenton oxidation of polycyclic aromatic hydrocarbons (PAHs) in soil pre-treated with Fenton reagents. Geology, Ecology, and Landscapes, 1(3): 197-204.

[41] Zhou, F., Chen, J. (2011): The Atlas of Microorganism and Macroinvertebrates in Freshwater. Second edition. – Chemistry Industry Press, Beijing.