Proceedings of the 2nd International Conference on Environmental Geoscience and Earth Ecology
DOI: 10.54254/2753-8818/37/20240213

Using the kriging method based on cross-validation stepwise variable selection to optimize anaerobic wastewater treatment process

Qilu Liu[1,2], Jingfang Shen[1]

[1]College of Informatics, Huazhong Agricultural University, Wuhan, 430070, China

[2]qilu_liu@webmail.hzau.edu.cn

Abstract. In this study, process parameters for wastewater treatment are optimised in order to improve the efficiency of energy recovery from wastewater treatment. In order to identify process parameters that have a significant impact on carbon recovery, we propose a stepwise forward variable selection method based on K-fold cross kriging to validate the loss ER.The results show that for complex wastewater treatment systems, the cross-validated kriging-based stepwise variable selection method can predict carbon recovery quickly and accurately. Finally, this paper combines the kriging prediction model with the PSO optimisation strategy to determine the optimal set points of process parameters with the objective of maximising carbon recovery.

Keywords: Anaerobic wastewater treatment, Kriging, Variable selection, K-fold cross-validation.

1. Introduction

UASB reactor is the most widely used anaerobic process for industrial wastewater treatment. However, the efficiency of UASB is not high enough and in order to maximise methane production, this paper wishes to optimise the process parameters for wastewater treatment. However, due to the large time delay in the wastewater treatment process, key parameters cannot be directly measured online [1]. In order to improve the optimisation efficiency of wastewater treatment parameters, surrogate models are usually used to predict the parameters that are difficult to be monitored online, and the surrogate models include artificial neural network [2], support vector machine [3] and Kriging method. Among them, Kriging is an accurate interpolation method, whose main advantage is that it not only provides the numerical response of the sample, but also the estimation error [4].

Considering the complexity of the wastewater treatment process, a kriging model is presented in Section 2.1 to ensure the accuracy of the predictions. In order to screen out process parameters that have a significant impact on carbon recovery efficiency, Section 2.2 presents a stepwise forward variable selection method based on K-fold cross-validation loss ER. Finally, Section 4 combines the prediction model with a particle swarm optimisation (PSO) algorithm to determine the optimal settings of the process parameters.

DOI: 10.54254/2753-8818/37/20240213

2. Methodology

2.1. Kriging

The Kriging model has been widely used to predict the response of real complex computer models. Usually, a set of training samples is first required to construct the Kriging model. The Kriging model is expressed as a sum of a polynomial and a Gaussian stochastic process:

$$\hat{y} = \mu(\boldsymbol{x}) + z(\boldsymbol{x}) \tag{1}$$

Where the $\mu(\boldsymbol{x})$ is the global estimate of the Kriging model, $z(\boldsymbol{x})$ denotes a Gaussian stochastic process providing local deviations. At the prediction point x, the prediction result $g(\mathbf{x})$ obeys a normal distribution: $g(\mathbf{x}) \sim N(\mu_g(\mathbf{x}), \sigma_g(\mathbf{x}))$, Assuming that the mean of stochastic process is 0, the covariance is defined by Eq.

$$E[z(\boldsymbol{x})z(\boldsymbol{w})] = \sigma^2 R(\boldsymbol{\theta}, \boldsymbol{w}, \boldsymbol{x}) \tag{2}$$

Where σ^2 denotes the process variance, $R(\boldsymbol{\theta}, \boldsymbol{w}, \boldsymbol{x})$ is the correlation function with parameters θ that determine the smoothness of the Kriging model. Maximum likelihood estimation (MLE) is often used to estimate regression coefficients $\boldsymbol{\beta}$, σ^2 and θ. The likelihood function is defined as the probability of N observations **y** given the parameters $\boldsymbol{\beta}$, σ^2, θ [5].

$$\mathrm{L}(\beta, \sigma^2, \theta | y) = p(y | \beta, \sigma^2, \theta) = \prod_{i=1}^{N} p(y_i | \beta, \sigma^2, \theta) \tag{3}$$

The estimates of each parameter can be obtained by maximizing the log-likelihood function:$\hat{\beta}, \hat{\sigma}^2, \hat{\theta} = \arg\max_{\beta,\sigma^2,\theta} \mathrm{L}(\beta, \sigma^2, \theta | \mathrm{y})$.

2.2. K-fold cross-validation stepwise variable selection method

Table 1. Symbols and definitions of influent and effluent quality parameters.

Parameter	Definition	Unite
Ss	Influent readily biodegradable substrate	g/m^3
Tss	Influent total suspended solids(total sediment solids)	g/m^3
Q	Influent flow	m^3/d
COD_0	Influent chemical oxygen demand	mg/L
COD_1	Effluent chemical oxygen demand	mg/L
COD removal efficiency	$(COD_0\text{-}COD_1)/COD_0$	/
PCH_4	Methane production	m^3
TN_0	Influent total nitrogen	mg/L
TP_0	Influent total phosphorus	mg/L
HRT	Hydraulic retention time	h
VL	Volume load	$COD/(m^3d)$
T	Water temperature	℃
V	Effective volume	m^3
H	Effective water depth	m
UFR	Upper flow rate	m/h

This paper proposes a stepwise forward variable selection method based on *K*-fold cross-validation loss *ER* (*K*=3) as shown in Figure 1. The approach follows the steps outlined below:

DOI: 10.54254/2753-8818/37/20240213

Step1: For a certain output parameter to be predicted, use the input parameter that has the largest P (Pearson correlation coefficient) with the output as the first selected input. The remaining input parameters are named set **S**.

Step2: Calculate the PCM for each remaining parameter in **S**. The PCM values are sorted in descending order, and the parameter with the largest PCM value is used as the candidate input variable. The PCM of the *a*-th input parameter is:

$$PCM_a = \mathbf{x_a}^{\mathrm{T}}\mathbf{x_a}\mathbf{x_a}^{\mathrm{T}}\mathbf{y} + \frac{1}{A}\sum_{j=1}^{A}\mathbf{x_a}^{\mathrm{T}}\mathbf{x_j}\mathbf{x_j}^{\mathrm{T}}\mathbf{y} \tag{4}$$

where $\mathbf{x}_a$ is the vector of the *a*-th input parameter, $\mathbf{x}_j$ is one of the input parameter vectors that have been selected, and *A* is the number of input parameters that have been selected. The calculation of PCM values evolved from the CME weights proposed by Sergio. The CME weights consider the correlation between input features, as well as the correlation between each input and output. The n-dimensional coefficient vector of CME is defined as follows: $W = C_{XX}C_{Xy}$,where {**X**, **y**} is the data. $\boldsymbol{C}_{XX}$, $\boldsymbol{C}_{Xy}$ denote the sample covariance matrices, $\boldsymbol{C}_{XX}=\mathbf{X}^{\mathrm{T}}\mathbf{X}/\mathrm{L}$ and $\boldsymbol{C}_{Xy}=\mathbf{X}^{\mathrm{T}}\mathbf{y}/\mathrm{L}$. L is the number of input parameters.

Step3: The candidate input variable is used as the next selected input variable to establish the Kriging and calculate the *ER*. Then move the selected input parameters out of **S**. In Kriging, the loss $l(\mathbf{x})$ at the prediction point **x** is defined as a squared form:

$$l(\mathbf{x}) = \left(f(\mathbf{x}) - f_K(\mathbf{x})\right)^2 \tag{5}$$

Where $f(\mathbf{x})$ is the true response and $f_K(\mathbf{x})$ is the Kriging prediction at the point **x**. The expectation $E[l(\mathbf{x})]$ of the $l(\mathbf{x})$ is used to represent the loss term of the Kriging at the prediction point **x**.

$$E[l(x)] = (E[f(x)] - E[f_K(x)])^2 + E[(f_K(x) - E[f_K(x)])^2] + E[(f(x) - E[f(x)])^2] \tag{6}$$

The first term of the above equation is the bias term; the second term is the predicted variance $\hat{\sigma}^2(\mathbf{x})$ of the Kriging model; and the third term represents the variance of the true response, which is the intrinsic noise of the true response data. The third term can be ignored when calculating the loss of the Kriging. The *ER* of the Kriging model at the prediction point **x** is calculated by Eq. (10).

$$ER(\mathbf{x}) = \underbrace{\left(y(\mathbf{x}) - \hat{\mu}(\mathbf{x})\right)^2}_{bias\ term} + \underbrace{\hat{\sigma}^2(\mathbf{x})}_{predicted\ variance} \tag{7}$$

Among them, the larger the deviation term, the greater the local prediction error of the Kriging prediction model at the prediction point x. The larger the variance term, the greater the global uncertainty of the prediction. Therefore, the ER is used to calculate the loss of the Kriging prediction model, and the local accuracy and global uncertainty of the Kriging prediction are fully considered.

For any point $\mathbf{x}_{iK}$ in the *K*-th fold sample set, assume that $\hat{y}^{-K}(\mathbf{x}_{iK})$ represents the predicted value of the Kriging model constructed using the remaining K-1 fold training samples at point $\mathbf{x}_{iK}$, and $y(\mathbf{x}_{iK})$ is the true response value at point $\mathbf{x}_{iK}$. The prediction bias of the Kriging prediction model at the point $\mathbf{x}_{iK}$ is:

$$e^2(\mathbf{x}_{iK}) = (y(\mathbf{x}_{iK}) - \hat{y}^{-K}(\mathbf{x}_{iK}))^2 \tag{8}$$

In summary, assuming that the value of K is 3, the K-fold cross-validation prediction variance and bias of the Kriging model are:

$$\sigma^2 = \frac{1}{3}\sum_{K=1}^{3}\left[\frac{1}{m}\sum_{i=1}^{m}\hat{\sigma}^2(\mathbf{x}_{iK})\right](K = 1, 2, 3) \tag{9}$$

$$e^2 = \frac{1}{3}\sum_{K=1}^{3}\left[\frac{1}{m}\sum_{i=1}^{m}e^2(\mathbf{x}_{iK})\right](K = 1, 2, 3) \tag{10}$$

where m is the number of samples in the K-fold training sample set. Combine the bias term with the variance term to obtain the *K*-fold cross-validation loss term *ER* (K=3).

DOI: 10.54254/2753-8818/37/20240213

$$ER = e^2 + \sigma^2 \tag{11}$$

$$\sigma^2 = \frac{1}{3}\sum_{K=1}^{3}\left[\frac{1}{m}\sum_{i=1}^{m}\hat{\sigma}^2(\mathbf{x}_{iK})\right](K=1,2,3) \tag{12}$$

$$e^2 = \frac{1}{3}\sum_{K=1}^{3}\left[\frac{1}{m}\sum_{i=1}^{m}e^2(\mathbf{x}_{iK})\right](K=1,2,3) \tag{13}$$

where m is the number of samples in the K-fold training sample set. Combine the bias term with the variance term to obtain the *K*-fold cross-validation loss term *ER* (K=3).

$$ER = e^2 + \sigma^2 \tag{14}$$

Step4: Repeat the above process until the *K*-fold cross-validation loss term *ER* meets the stop criterion (|GE|<0.05 or GE>0). The GE is calculated as follows:

$$GE = (ER^{i} - ER^{i-1})/ER^{ini} \tag{15}$$

Where ER^{ini} is the loss for building a model using only the first input variable, ER^{i} is the loss for building a prediction model using *i* input variables, ER^{i-1} is the loss for building a prediction model using *i-1* input variables.

Step5: Construct the Kriging prediction model by using the selected *i*-1 input variables.

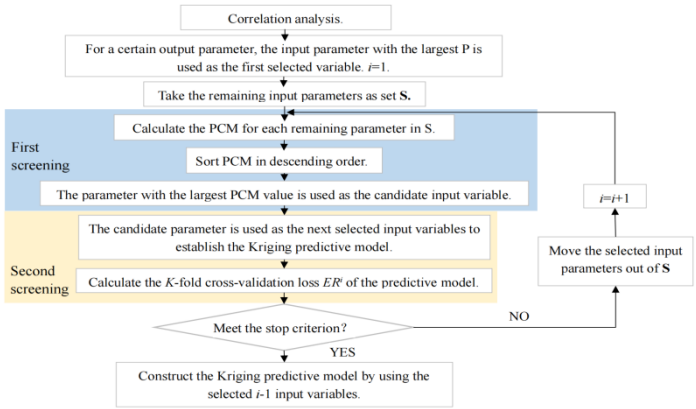

Figure 1. The process of variable selection.

3. Results

3.1. Results of the variable selection

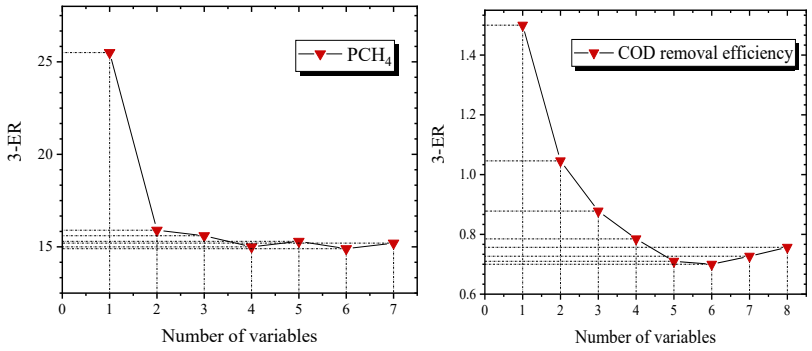

Figure 2. The process of forward variable selection based on 3-fold cross-validation loss.

DOI: 10.54254/2753-8818/37/20240213

Fig. 2 shows the process of variable selection based on calculating the cross-verification loss 3-ER. The addition of the first few variables makes the cross-verification loss *ER* decrease rapidly. However, as the number of variables increases, the curve of the 3-fold cross-validation loss *ER* gradually flattens or rises. The change law of the loss term *ER* is determined by the change law of the deviation term and the variance term. Too many dimensions will instead increase the uncertainty of prediction and increase the prediction variance.

Table 2. The result of 3-fold cross-validation forward stepwise variable selection.

Effluent variables	Selected input parameters	Number
PCH_4	Q,COD removal efficiency	2
COD removal efficiency	VL, HRT, COD_0, T, UFR	5

Table 2 shows the variables selected to predict methane production and COD removal efficiency. The variable selection results showed that influent flow and COD removal rate significantly affected methane production. Maximizing COD removal also maximizes methane production. In summary, we can optimize the COD removal rate by adjusting *HRT*, COD_0, *T*, and *UFR*.

3.2. Model prediction

Modelling accuracy calculation methods include: mean absolute geometric error ($MAGE = \exp\left(\frac{1}{N}\sum_{j=1}^{N}\log(y_j/\hat{y}_j)\right)$)mean absolute error ($MAE = \frac{1}{N}\left(\sum_{j=1}^{N}|y_j-\hat{y}_j|\right)$), root mean square error ($RMSE = \sqrt{\frac{\sum_{j=1}^{N}(y_j-\hat{y}_j)^2}{N}}$) and Mean Absolute Percentage Error (MAPE). where N is the number of sample points in the test set, $\hat{y}_j$ is the model predicted response, y_j is the expensive simulated response (true response).

3.2.1. Prediction accuracy of adaptive weighted average Kriging. This section measures the prediction effectiveness of the prediction model for COD removal efficiency by calculating the absolute scale error and the relative scale error. Only the key input parameters selected in Table 2 are used to build the Kriging predictive model.

Table 3. The performance of the predictive model on the test set.

Effluent quality variable	MAE	RMSE	MAGE
PCH_4	0.54	0.65	1.0026
COD removal efficiency	0.0910	0.1279	0.9009

Based on Table 3, It can be concluded that the MAE and RMSE of the prediction model are small enough. In addition, the established model has outstanding performance in terms of relative indicators. The MAGE of the PCH_4 and COD removal efficiency are very close to 1 , indicating that the goodness-of-fit of the predictive model is high.

4. Optimization results and analysis

Through the verification in section 3, the Kriging model can replace the actual monitoring of wastewater quality and achieve optimal design. A prediction model was established for COD removal efficiency, and then optimization algorithms were used to optimize the set points of *HRT*, COD_0, *T* and *UFR*. The optimization model can be expressed mathematically as:$Max\ f_{COD\ \text{removal}}(x_1, x_2, x_3, x_4)$,Among them, $f_{COD\ \text{removal}}(x_1, x_2, x_3, x_4)$ represents the prediction function of COD removal efficiency, x_1, x_2, x_3, x_4 are optimal set points for *HRT*, COD_0, *T* and *UFR*, respectively. The set value of controllable parameters is obtained by the PSO algorithm. The optimized parameters are shown in Table 3.

DOI: 10.54254/2753-8818/37/20240213

Table 4. Values of design variables.

Objective function	Design variable	Range	Optimized value	COD removal efficiency(%)	$PCH_4(m^3)$
ax $f_{COD\ removal}$ (x_1, x_2, x_3, x_4)	HRT (h)	[5,80]	39.5	73.4	596.24
	COD_0 (mg/L)	[1000, 15000]	6769.34		
	UFR (m/h)	[0, 2]	0.453		
	T (℃)	[20,40]	29.5		

The optimization time based on the Kriging model is short. In contrast, if the online monitoring of sewage COD concentration is directly used, it is not only extremely time-costly, but also has a lag. Therefore, the optimization efficiency is greatly improved by using the Kriging prediction model based on K-fold cross-validation stepwise variable selection method and PSO optimization method.

5. Conclusions

In this study, the carbon source recovery efficiency of anaerobic sewage treatment system was optimized by combining the Kriging prediction model based on K-fold cross-validation stepwise variable selection method and PSO optimization strategy. The results showed that the COD removal efficiency was closely related to PCH_4, and *HRT*, COD_0, *T* and *UFR* could significantly affect the COD removal efficiency.

Therefore, the goal of this study is to maximize COD removal efficiency, combined with the prediction model of COD removal efficiency, to find the optimal set points of *HRT*, COD_0, *T* and *UFR*.According to the optimization results, the control parameters of the sewage treatment system can be modified to improve the efficiency of energy recovery. This optimization scheme provides effective parameter design and solutions for other optimization problems in the field of wastewater treatment.

Acknowledgements

We thank Fundamental Research Funds for the Central Universities (No. 2662022LXYJ003).

References

[1] Zounemat-Kermani M, Alizamir M, Keshtegar B, Batelaan O, Hinkelmann R. Prediction of effluent arsenic concentration of wastewater treatment plants using machine learning and kriging-based models. Environ Sci Pollut Res Int. 2022 Mar;29(14):20556-20570.

[2] Jorge E Hurtado, Diego A Alvarez. (2001). Neural-network-based reliability analysis: a comparative study, Computer Methods in Applied Mechanics and Engineering, Volume 191, Issues 1–2.

[3] Basudhar, A., & Missoum, S. (2008). Adaptive explicit decision functions for probabilistic design and optimization using support vector machines. Computers and Structures, 86(19–20), 1904–1917.

[4] B. J. Bichon, M. S. Eldred, L. P. Swiler, S. Mahadevan and J. M. McFarland. (2012). Efficient Global Reliability Analysis for Nonlinear Implicit Performance Functions[J]. AIAA Journal, 46(10) : 2459-2468.

[5] K, V, MARDIA, et al. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression[J]. Biometrika, 71(1):135-146.