Int. J. Electrochem. Sci., 16 **(2021)** Article ID: 210253, doi: 10.20964/2021.02.21

International Journal of
ELECTROCHEMICAL
SCIENCE
www.electrochemsci.org

Stray Current Prediction Model for Buried Gas Pipelines Based on Multiple Regression Models and Extreme Learning Machine

Jiansan Li[1], *Zhenbin Liu*[1,*], *Hong Yi*[2], *Guiyun Liu*[2] *and Yifan Tian*[1]

[1] School of Mechanical and Automotive Engineering, South China University of Technology, Guangzhou 510640, PR China
[2] Guangzhou Gas Group Co., Ltd., Guangzhou 510635, PR China
[*]E-mail: mezbliu@mail.scut.edu.cn

Received: 3 October 2020 / *Accepted:* 25 October 2020 / *Published:* 31 December 2020

Serious stray current corrosion poses a threat to the sustainable and safe use of buried gas pipelines. To exactly predict the stray current of buried gas pipelines and take timely action to reduce stray current corrosion on buried pipelines, the multiple linear regression (MLR) model, multiple nonlinear regression (MNLR) model, extreme learning machine (ELM) model and extreme learning machine processed by principal component analysis (PCA-ELM) model are established in this work. The stray current data obtained on site are applied to establish the above four prediction models. The predicted results suggest that the neural network models perform better at prediction than the traditional multiple regression models, and the proposed PCA-ELM model yields the smallest prediction errors, leading to a higher prediction accuracy and better generalization performance than the other three prediction models. However, the activation function and the number of hidden layer nodes in the neural network models should be selected and tested carefully. With the local optimization method, the proposed PCA-ELM model prefers the sine activation function and 18 hidden layer nodes. In summary, the proposed PCA-ELM model can be used for stray current prediction of buried gas pipelines or in other prediction studies.

Keywords: Multiple regression model; Extreme learning machine; Principal component analysis; Stray current; Prediction

1. INTRODUCTION

With the development of urban rail transit systems and high-voltage power transmission and transformation systems, stray current corrosion in buried gas pipelines used for urban gas transmission and distribution systems is becoming increasingly prominent [1-4]. In particular, pipelines with defects in their external anti-corrosion coating are easily affected by stray current corrosion [5], which can lead to the perforation of the pipelines and leakage of the internal medium, resulting in huge economic

losses and serious environmental pollution [6]. In addition, stray current corrosion of buried gas pipelines is a type of electrochemical corrosion. Therefore, by studying the relevant factors that affect the stray current of buried gas pipelines and establishing a prediction model between the stray current and influencing factors, we can take the necessary measures in time to reduce the electrochemical corrosion of buried pipelines, which is of great significance for the safe operation of pipelines.

Recently, mathematical regression models and neural network models have been applied in various engineering fields for predictions. Thoe [7] established multiple linear regression (MLR) models to forecast the daily water quality of Hong Kong beaches. More [8] established a MLR model and multiple nonlinear regression (MNLR) model to determine the chromium removal efficiency (CRE) in the cathode chamber of Bioelectrochemical system, concluding that the MLR model was better suited than the MNLR model for predicting chromium removal behavior. With the development of artificial neural networks, Rezaeianzadeh [9] used an artificial neural network (ANN), adaptive neuro-fuzzy inference system (ANFIS), MLR model and MNLR model to forecast maximum daily flow and proposed the MNLR model as a simple way to predict the maximum daily flow. Cao [10] used the BP neural network model to predict the stray current of buried gas pipelines, and Wang [11] established a quantum particle swarm optimization neural network (QPSO-NN) model to predict the stray current density; both studies obtained a good prediction performance. As a type of single hidden layer feedforward neural network, the extreme learning machine (ELM) has been widely used in various fields, especially in the fields of classification [12-16] and prediction [17-25]. After verifying by the simulation results, Huang [26] concluded that ELM achieved a better generalization performance for regressions and achieved a much faster learning speed (up to thousands of times faster) than traditional support vector machine (SVM) and least squares support vector machine (LS-SVM).

In this paper, we propose an approach for training ELM networks based on principal component analysis (PCA). PCA is widely used for data analysis and dimension reduction in applications throughout science and engineering [27-29]. The trained PCA-ELM model is applied to predict the stray current of buried gas pipelines, and the predicted results are compared with those from models established by traditional multiple linear regression and multiple nonlinear regression. The prediction accuracies of the models are analysed and compared, and the results provide references for methods and a certain theoretical basis for stray current prediction of buried gas pipelines.

2. EXPERIMENTAL

2.1. Multiple linear regression model

Multiple linear regression (MLR) is a linear regression relationship in which the dependent variable is affected by two or more independent variables. MLR is a commonly used method in mathematical statistics. The basic principle of MLR is similar to that of the one-variable linear regression model. Setting the dependent variable as *y* and the independent variable as x_i (i=1, 2, ..., k), the basic MLR model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \quad (1)$$

where $\beta_0,\beta_1,...,\beta_k$ are k+1 unknown parameters; β_0 is the regression constant; $\beta_1,\beta_2,...,\beta_k$ are the regression coefficients; y is the explained variable; $x_1,x_2,...,x_k$ are k explanatory variables that can be precisely controlled; and ε is random error.

If there are n sets of sample data, denoted as (x_{i1}, x_{i2}, …, x_{ik}, y_i), (where i=1, 2, ..., n), then n regression equations can be established:

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_k x_{1k} \\ y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_k x_{2k} \\ \cdots\cdots \\ y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_k x_{nk} \end{cases} \quad (2)$$

Written in matrix form:

$$Y=X\beta \quad (3)$$

where $Y = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix}$, $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdots \\ \beta_k \end{bmatrix}$ and $X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$.

The estimated value $\hat{\beta}$ of the regression coefficient β in the MLR model can be obtained by the least squares method:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad (4)$$

from which we can calculate that $\hat{\beta} = \left[\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,...,\hat{\beta}_k\right]^T$. Then, the MLR model obtained is as follows:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k \quad (5)$$

2.2. Multiple nonlinear regression model

Multiple nonlinear regression (MNLR) is a nonlinear regression in which the dependent variable is affected by two or more independent variables. This paper uses SPSS (version 23.0) and 1stOpt (version 15.0) to establish and analyse the MNLR model [30]. First, the optimal unitary nonlinear regression model of y on each independent variable x_i is established in an exploratory way. All the curve models in SPSS (version 23.0) are selected to fit the training set, and the optimal unitary curve model is selected by comparing the coefficient of determination (R^2). The larger the coefficient of determination, the better the model fitting effect. Second, all the optimal unitary curve models are artificially synthesized into a multiple nonlinear model. Then, the parameters of the multiple nonlinear model are estimated by using 1stOpt (version 15.0), and finally, the MNLR model is acquired. The process for establishing the MNLR model is shown in Figure 1.

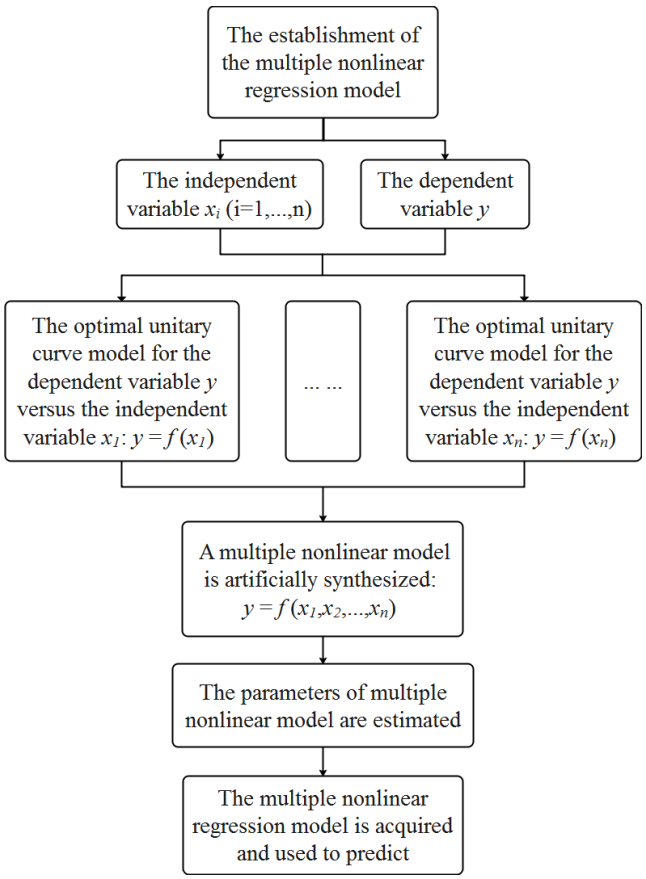

Figure 1. Process for establishing the MNLR model

2.3. Principal component analysis

Principal component analysis (PCA) is a data dimension reduction method in statistics. PCA transforms the original multiple correlated variables into a few unrelated principal components, and each principal component is a linear combination of the original variables. The selected principal components should retain most of the information of the original variables as much as possible to reduce the data dimension. PCA can eliminate the linear correlation between variables, eliminate the redundancy of data, reduce the dimension of the model input variables, simplify the structure of networks, and improve the model training and prediction speed.

Suppose X_{nm} is a matrix composed of n samples and each sample is m-dimensional. The matrix form is written as follows:

$$X_{nm}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \tag{6}$$

Principal component analysis (PCA) can be divided into the following five steps:

1) Standardizing the original data. The original data are standardized according to Eq. (7), and the standardized matrix X_{nm}^{*} is obtained.

$$x_{ij}^{*}=\frac{x_{ij}-\overline{x}_j}{s_j} \tag{7}$$

where $\overline{x}_j=\frac{1}{n}\sum_{i=1}^{n}x_{ij}$, $s_j=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij}-\overline{x}_j\right)^2}$.

2) Establishing the correlation coefficient matrix. The correlation coefficient is calculated by the standardized matrix X_{nm}^{*} according to Eq. (8), and the correlation coefficient matrix R_{nm} is obtained.

$$r_{ij}=\frac{\sum_{k=1}^{n}\left(x_{ki}^{*}-\overline{x}_i^{*}\right)\left(x_{kj}^{*}-\overline{x}_j^{*}\right)}{\sqrt{\sum_{k=1}^{n}\left(x_{ki}^{*}-\overline{x}_i^{*}\right)^2\sum_{k=1}^{n}\left(x_{kj}^{*}-\overline{x}_j^{*}\right)^2}} \tag{8}$$

3) Solving the eigenvalues and eigenvectors of the correlation coefficient matrix. The formula $\left|\lambda E-R\right|=0$ is used to solve the eigenvalue $\lambda_i\left(i=1,2,...,m\right)$, and the eigenvalues are arranged in descending order: $\lambda_1\geq\lambda_2\geq...\geq\lambda_m\geq 0$. Then, the formula $\left(\lambda_i E-R\right)A=0$ is used to solve the eigenvector $A_i=\left[a_{i1},a_{i2},...,a_{im}\right]^T$ of the corresponding eigenvalue λ_i.

4) Calculating the explained variance and cumulative variance of each principal component. The explained variance of the *i*th principal component is: $G_i=\lambda_i/\left(\sum_{j=1}^{m}\lambda_j\right)$, where $i=1,2,...,m$. The cumulative variance of the first *l* principal components is: $G(l)=\sum_{i=1}^{l}G_i$. If $G(l)$ is more than 80%, then the first *l* principal components are taken as the input variables of the networks; thus, the input dimension of the networks is reduced from *m* to *l*.

5) Calculating the principal component matrix. The principal component matrix P_{nl} composed of *n* samples corresponding to *l* principal components is obtained as follows:

$$P_{nl}=X_{nm}^{*}A_{ml} \tag{9}$$

where X_{nm}^{*} is the standardized matrix and $A_{ml}=\left[A_1,A_2,...,A_l\right]=\begin{bmatrix}a_{11} & a_{21} & \cdots & a_{l1}\\ a_{12} & a_{22} & \cdots & a_{l2}\\ \vdots & \vdots & \vdots & \vdots\\ a_{1m} & a_{2m} & \cdots & a_{lm}\end{bmatrix}$.

2.4. Extreme learning machine

Extreme learning machine (ELM) is a new type of feedforward neural network with a single hidden layer [31]. ELM replaces the iterative process of traditional parameter optimization by solving linear equations. The solution obtained is taken as the weight of the output layer network so that the training of the network can be completed at one time without iteration. In addition, in the training process of ELM, the input weight matrix (w) and the hidden layer biases (b) are randomly generated without adjustment, which simplifies the parameter selection of the algorithm. This process also overcomes the shortcomings of traditional feedforward neural networks, such as the complex training parameters, slow training speed and ease of falling into the local minima [32-33].

Assume N arbitrarily different training samples $\{(x_i,t_i)\}_{i=1}^{N}$, where $x_i=[x_{i1},x_{i2},...,x_{in}]^T\in R^n$ is the n-dimensional input vector of the ith sample, and $t_i=[t_{i1},t_{i2},...,t_{im}]^T\in R^m$ is the corresponding expected output vector. Assuming that the number of hidden layer nodes is L and the activation function is $g(x)$, then the ELM model can be written as follows:

$$H\beta=T \tag{10}$$

where $H=\left[h(x_1)^T,h(x_2)^T,...,h(x_N)^T\right]^T=\begin{bmatrix} g(w_1x_1+b_1) & \cdots & g(w_Lx_1+b_L) \\ \vdots & \ddots & \vdots \\ g(w_1x_N+b_1) & \cdots & g(w_Lx_N+b_L) \end{bmatrix}_{N\times L}$.

Here, $w_i=[w_{i1},w_{i2},...,w_{in}]^T$ is the weight between the input layer and the ith hidden layer node, b_i is the bias of the ith hidden layer node, $\beta=[\beta_1,\beta_2,...,\beta_L]^T$ is the output weight matrix connecting the hidden layer nodes and the output layer, and $T=[t_1,t_2,...,t_N]^T$ is the expected output sample matrix. The input weight matrix (w) and the hidden layer biases (b) are randomly generated. To minimize the training error, the output weight matrix (β) in Eq. (10) can be obtained by calculating the least squares solution as follows:

$$\hat{\beta}=H^{+}T \tag{11}$$

where H^{+} is the Moore–Penrose generalized inverse of the output matrix of hidden layer H.

In summary, the basic steps of ELM are as follows [34]:

a) Given N arbitrarily different training samples $\{(x_i,t_i)\}_{i=1}^{N}$, set the number of hidden layer nodes and the activation function;

b) Randomly set the input weight matrix w and the hidden layer biases b;

c) Calculate the output matrix of hidden layer H;

d) Calculate the output weight matrix $\hat{\beta}$ using Eq. (11).

2.5. The indicators of the prediction accuracy

In this study, the mean absolute error (MAE), mean absolute percentage error (MAPE) and root mean square error (RMSE) are used as the indicators of the prediction accuracy of the model. The smaller the above error values are, the better the prediction accuracy. The equations are expressed as follows:

$$MAE=\frac{1}{N}\sum_{i=1}^{N}|y_i-\hat{y}_i| \tag{12}$$

$$MAPE=100\times\frac{1}{N}\sum_{i=1}^{N}\frac{|y_i-\hat{y}_i|}{y_i} \tag{13}$$

$$RMSE=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2} \tag{14}$$

where y_i and $\hat{y}_i$ represent the measured and predicted values of the ith sample in a testing set, respectively; N is the number of samples in a testing set.

2.6. The measurement of data

To predict the stray current of buried gas pipelines in Guangzhou, Guangdong Province, China, 20 points along the high-pressure pipelines were selected. At each testing point, we measured the pipe-to-soil potential, the soil moisture content, the soil resistivity, the buried depth of the pipeline, the pH of the soil and the stray current. Among the measurements, the measurement process of the pipe-to-soil potential included a testing pit to be dug at a depth of 0.5m directly above the pipeline and a reference electrode to be placed. A small test piece, which was made of the same material as the pipeline and set close to the reference electrode to eliminate the IR drop as much as possible, was connected to the pipeline through the test pile. A multimetre was used to measure the potential between the small test piece and reference electrode for 1 hour, and the average value was taken as the pipe-to-soil potential. The soil moisture content and pH were measured as follows: according to "GB/T 19285-2014 Corrosion protection engineering inspection of buried steel pipeline" (hereinafter referred to as "Standard"), five soil samples were randomly selected from the testing pit for analysis, and the average values of the analysis results were taken as the soil moisture content and pH of the testing point. The soil resistivity was measured as follows: according to the "Standard", the grounding resistance measurement instrument was applied and the equidistant four-electrode method was used to measure soil resistivity 5 times continuously, and the average value was taken as the soil resistivity of the testing point. The buried depth of the pipeline was measured using an RD8100 pipeline detector. As shown in Figure 2, after measuring the pipeline current of each testing point and the current provided by the cathodic protection power supply, the absolute value of the subtraction of these two currents was taken as the stray current of the corresponding testing point.

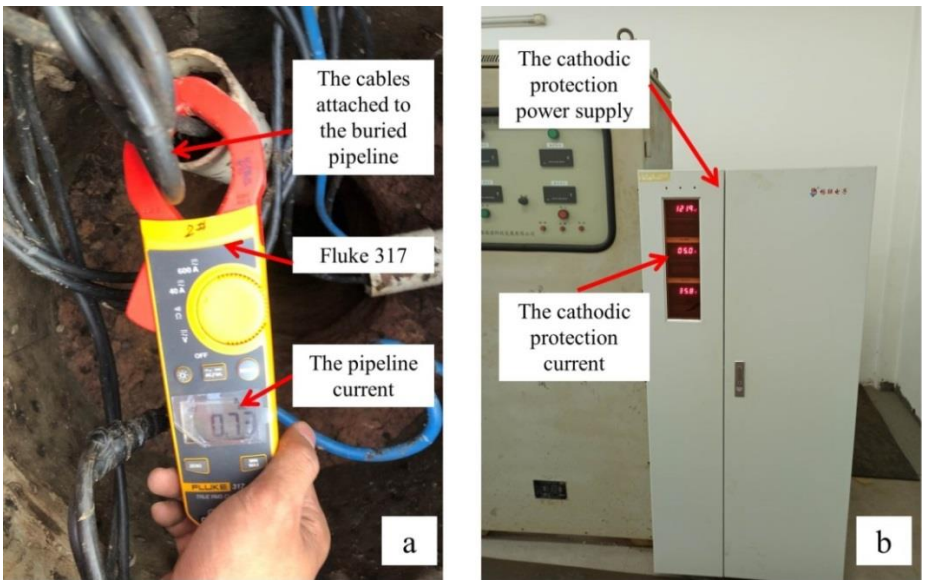

Figure 2. Measurement of (a) the pipeline current and (b) the cathodic protection current

Using the above measurement methods, 20 samples were obtained, as shown in Table 1. Fifteen samples (Nos. 2, 4~7, 9~14, 16~17, and 19~20) were randomly selected from Table 1 as the training set of each prediction model for establishment, and the remaining 5 samples (Nos. 1, 3, 8, 15, and 18) were selected as the testing set of each prediction model to test the prediction performance.

To simplify the expression, y is used to represent the stray current of buried gas pipelines and x_1, x_2, x_3, x_4, x_5 are used to represent the pipe-to-soil potential, the soil moisture content, the soil resistivity, the buried depth of pipeline and the pH of the soil, respectively.

Table 1. Measurement data

No.	The pipe-to-soil potential /V	The soil moisture content /%	The soil resistivity /(Ω·m)	The buried depth of pipeline /m	The pH of the soil	The stray current /A
1	-1.26	22.3	179.7	1.5	7.1	2.27
2	-1.11	19	473.7	5.3	7.5	1.02
3	-1.75	25	279.7	1.25	7.5	3.07
4	-1.73	24.7	396.7	1.46	7.1	1.35
5	-1.32	16.7	822.3	1.8	8.1	1.65
6	-2.94	14.7	1039	3.7	6.8	2.40
7	-1.35	27.7	104.7	1.88	7.9	2.32
8	-1.82	18	162.8	1.4	7.1	2.25
9	-1.42	18	393	1.5	6.7	2.03
10	-1.37	19.7	232.7	2.74	6.7	1.95
11	-1.87	16.3	1204	1.6	6.6	1.87
12	-1.21	23	305	1.5	7.4	2.33
13	-0.97	18	1255.3	1.8	8	0.72
14	-3.15	15	1655.7	3.7	7.6	2.21
15	-1.37	20.3	224.7	1.88	6.4	2.19
16	-2.72	25	100.1	1.4	6.5	3.92
17	-1.31	21	421.8	1.5	6.3	2.25
18	-1.61	24.7	225.4	2.74	6.1	2.64
19	-1.93	17	1004	1.6	6.1	2.07
20	-1.38	25.7	261.7	1.5	6.8	2.45

3. RESULTS AND DISCUSSION

3.1. The MLR model

In an MLR model, if there is a linear correlation between the independent variables, then there will be an increase in the standard error, resulting in a decrease of the prediction accuracy of the model [35]. To eliminate the influence of multicollinearity, this paper utilizes the training set and the stepwise regression method of SPSS (version 23.0) to establish the MLR model of the stray current of buried gas pipelines [36]. The fitting results and coefficient analysis are separately shown in Table 2 and Table 3.

Table 2. Fitting results of the MLR model

Model	Correlation coefficient R	Coefficient of determination R^2	Adjusted R^2	F	Significance
1	0.549	0.301	0.247	5.603	0.034
2	0.842	0.709	0.661	14.636	0.001
3	0.894	0.799	0.744	14.546	0.000

It can be seen from Table 2 that the correlation coefficient and coefficient of determination of Model 3 are the largest of all the models, indicating that Model 3 fits the training set best. In addition, the correlation coefficient of Model 3 is 0.894, meaning that the independent variables in Model 3 have a high correlation with the dependent variable [37]. Furthermore, the significance of the t-test is less than 0.05, which indicates that Model 3 is statistically significant. In Table 3, the variance inflation factor (VIF) of each variable in Model 3 is less than 4, indicating that there is no collinear error among the independent variables. The significance of the t-test for each variable in Model 3 is less than 0.05, which indicates that the independent variables in Model 3 are significant and the fitting is significant [38]. However, based on the stepwise regression method, the independent variables x_2 and x_5 did not enter Model 3. In statistical theory, this indicates that the soil moisture content and the pH of soil have no significant influence on the stray current of buried gas pipelines [39]. According to the coefficient analysis results of Model 3 presented in Table 3, the MLR model of the stray current is obtained as follows:

$$\hat{y} = 1.471 - 0.946x_1 - 0.001x_3 - 0.199x_4 \quad (15)$$

Table 3. Coefficient analysis of the MLR model

Model	Variate	Unstandardized coefficients		t	Significance	VIF
		B	Standard error			
1	Constant	1.029	0.456	2.255	0.042	
	x_1	-0.586	0.248	-2.367	0.034	1.000
2	Constant	1.180	0.308	3.824	0.002	
	x_1	-0.899	0.183	-4.916	0.000	1.210
	x_3	-0.001	0.000	-4.104	0.001	1.210
3	Constant	1.471	0.299	4.924	0.000	
	x_1	-0.946	0.160	-5.899	0.000	1.232
	x_3	-0.001	0.000	-4.186	0.002	1.261
	x_4	-0.199	0.090	-2.210	0.049	1.099

The prediction results of using Eq. (15) to predict the testing set are given in Figure 3. It can be seen from Figure 3 that the predicted values are not very close to the measured values. After calculation, the coefficient of determination of the predicted values is only 0.091, which is far smaller

than that of the training set (R^2=0.799). Therefore, the generalization performance of the MLR model is not satisfactory, which may lead to the poor prediction accuracy [40-41].

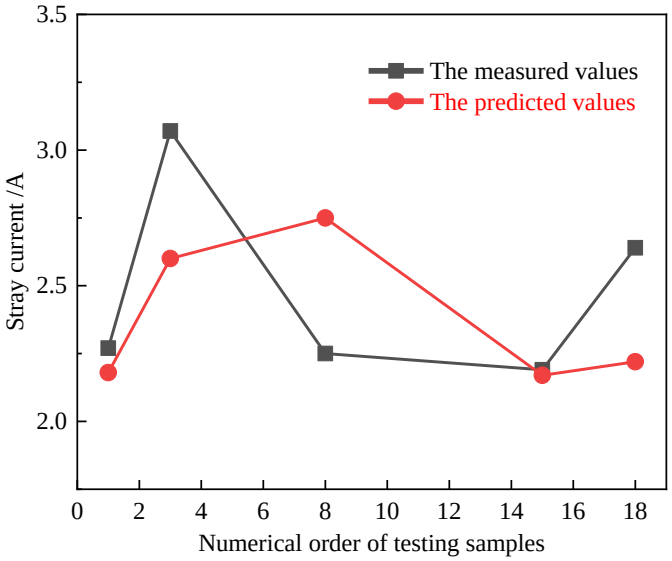

Figure 3. Predicted results of the MLR model

3.2. The MNLR model

Table 4. Statistical results of each model during the fitting on x_1

Curve models	Statistical summary			The estimation of the parameters			
	R^2	F	Significance	Constant	Parameter 1	Parameter 2	Parameter 3
Linear	0.301	5.603	0.034	1.029	-0.586		
Logarithmic curve	——	——	——	——	——	——	——
Inverse function curve	0.376	7.822	0.015	3.448	2.154		
Quadratic curve	0.369	3.513	0.063	-0.965	-2.798	-0.533	
Cubic curve	0.373	2.179	0.148	0.400	-0.438	0.738	0.212
Compound curve	0.280	5.051	0.043	1.118	0.734		
Power function curve	——	——	——	——	——	——	——
S-curve	0.416	9.247	0.009	1.457	1.241		
Growth curve	0.280	5.051	0.043	0.112	-0.309		
Exponential curve	0.280	5.051	0.043	1.118	-0.309		
Logistic curve	0.280	5.051	0.043	0.894	1.363		

In this paper, SPSS (version 23.0) and 1stOpt (version 15.0) are applied to establish the MNLR model of the stray current of buried gas pipelines. First, the optimal unitary nonlinear regression model

of y on each independent variable x_i (i=1,...,5) is established using SPSS (version 23.0). The independent variable x_1 is taken as an example. All the curve models in SPSS (version 23.0) are selected to perform curve fitting on the training set, and the statistical results of each model during the fitting on x_1 are obtained in Table 4.

Since the values of the independent variable x_1 are all negative, it is impossible to establish a logarithmic model and power function model for the dependent variable y to the independent variable x_1. Comparing the coefficient of determination (R^2) of each curve model presented in Table 4, we find that the R^2 of the S-curve model is the largest and the significance of the t-test is less than 0.05, which indicate that the S-curve model is statistically significant. Therefore, the S-curve model is selected for fitting, and the optimal unitary curve model of the dependent variable y to the independent variable x_1 is:

$$y = e^{1.457+\frac{1.241}{x_1}} \tag{16}$$

Using the same method, the optimal unitary curve models of the dependent variable y to the independent variables x_2, x_3, x_4, x_5 are established, and the results are as follows:

$$y = 6.973 - 0.549x_2 + 0.015{x_2}^2 \tag{17}$$

$$y = 1.557 + \frac{156.505}{x_3} \tag{18}$$

$$y = 11.853 - 11.074x_4 + 3.694{x_4}^2 - 0.376{x_4}^3 \tag{19}$$

$$y = 17.766 \times e^{-0.316x_5} \tag{20}$$

Second, the above optimal unitary curve models (16) ~ (20) are artificially synthesized into a multiple nonlinear model as follows:

$$y = \beta_0 + e^{\beta_1+\frac{\beta_2}{x_1}} + \beta_3 x_2 + \beta_4 {x_2}^2 + \frac{\beta_5}{x_3} + \beta_6 x_4 + \beta_7 {x_4}^2 + \beta_8 {x_4}^3 + \beta_9 e^{\beta_{10}x_5} \tag{21}$$

Then, the parameters of Eq. (21) are estimated by 1stOpt (version 15.0), and the results are presented in Table 5. In Table 5, the correlation coefficient (R) of the MNLR model is 0.939, meaning that the independent variables in the MNLR model have very high correlation with the dependent variable [37]. In addition, the coefficient of determination (R^2) is 0.882, which is higher than that of the MLR model (R^2=0.799). Therefore, the MNLR model fits the training set better than the MLR model. Finally, the MNLR model of the stray current of buried gas pipelines is obtained as follows:

$$\hat{y} = 0.0002 + e^{2.270+\frac{0.030}{x_1}} - 0.032x_2 + 0.001{x_2}^2 + \frac{147.602}{x_3} - 9.313x_4 + 3.218{x_4}^2 - 0.333{x_4}^3 + 55890.275e^{-1.827x_5} \tag{22}$$

Table 5. Parameter estimation of the MNLR model

No.	Projects calculated	Results obtained	Parameters	Estimated results
1	Algorithm	Levenberg-Marquardt	β_0	0.000198
2	Iterations	79	β_1	2.270343
3	RMSE	0.243531	β_2	0.030144

4	SSE	0.889612	β_3	-0.032231
5	R	0.938958	β_4	0.001071
6	R^2	0.881643	β_5	147.602363
7	Chi-Squared	0.284817	β_6	-9.313477
8	F-Statistic	96.837368	β_7	3.218090
9			β_8	-0.333109
10			β_9	55890.274765
11			β_{10}	-1.826901

Eq. (22) is applied to predict the testing set, and the results are shown in Figure 4. From Figure 4, the predicted values slightly fluctuate compared with the measured values, which is similar to the predicted results of the MLR model. After calculation, the coefficient of determination of the predicted values is 0.368, which is far less than that of the training set (R^2=0.882). Therefore, similar to the MLR model, the MNLR model also has poor generalization, which may affect the prediction accuracy of the model [40-41].

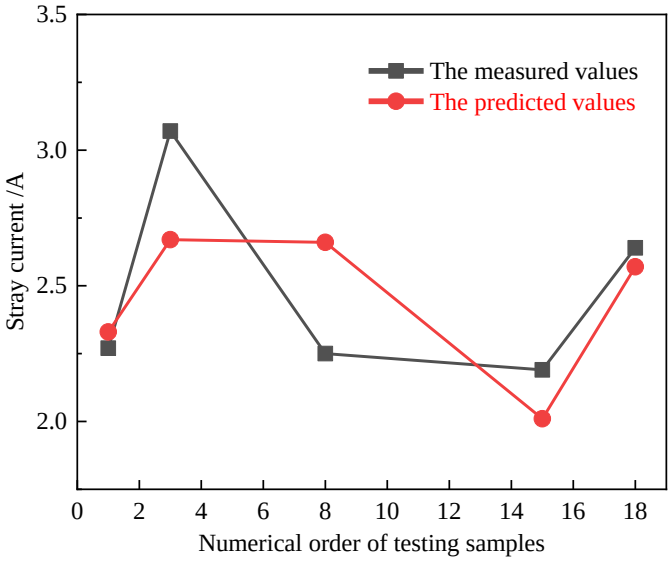

Figure 4. Predicted results of the MNLR model

3.3. PCA

In this study, PCA is applied to preprocess the original data into a set of linearly uncorrelated variables, which are called principal components. This technique reduces the dimension of the data, which reduces the computational memory and time. Table 6 shows the explained variance and cumulative variance of each principal component. It can be seen from Table 6 that the cumulative variance of the first three principal components is 83.410%, which explains 83.410% of the total variance and contains the acceptable information of the total variance. Therefore, the first three principal components are extracted as evaluation indexes and are taken as the new input variables of the networks, meaning that the input dimension of the networks has been reduced from 5 to 3. After

extracting the principal components, the new dataset processed by PCA is obtained, as shown in Table 7. Later, we will utilize the data in Table 7 to train the ELM model and apply the trained ELM model to predict the stray current of buried gas pipelines.

Table 6. Explained variance and cumulative variance of each principal component

Principal components	Explained variance/%	Cumulative variance/%
1	44.947	44.947
2	23.238	68.185
3	15.224	83.410
4	12.759	96.168
5	3.832	100.000

Table 7. New dataset processed by PCA

No.	Principal component 1	Principal component 2	Principal component 3
1	-1.19688	0.30228	-0.14317
2	1.07341	1.62349	2.54641
3	-1.17087	0.42684	-0.29740
4	-0.97066	-0.04661	-0.16813
5	0.75729	1.79345	-0.93310
6	2.91109	-0.99909	0.71549
7	-1.74951	1.37360	0.45119
8	-0.26685	-0.19166	-0.44203
9	-0.25391	-0.34043	-0.50868
10	-0.26908	-0.13659	0.77577
11	1.36225	-0.66866	-1.14454
12	-1.12201	0.78205	-0.25644
13	0.88611	2.04275	-1.21584
14	3.92938	0.05077	0.10937
15	-0.74030	-0.68164	0.14276
16	-0.86664	-1.76572	0.21993
17	-0.78542	-0.78465	-0.28409
18	-0.93918	-1.15102	1.21630
19	0.96305	-1.44384	-0.84730
20	-1.55125	-0.18533	0.06353

3.4. The ELM model

This paper uses MATLAB (version R2018) to establish the ELM neural network model to predict the stray current of buried gas pipelines. Five factors that affect the stray current of buried gas

pipelines (the pipe-to-soil potential, the soil moisture content, the soil resistivity, the buried depth of pipeline and the pH of the soil) are taken as the input parameters of the ELM model, and the stray current of buried gas pipelines is taken as the output parameter.

The activation function and the number of hidden layer nodes have strong influences on the prediction accuracy of ELM neural networks [42]. If the number of hidden layer nodes is too small, then the ELM networks cannot learn well and the prediction error will be large. If the number of hidden layer nodes is too large, then the training time of networks will increase and the phenomenon of overfitting is prone to occur. According to the *Kolmogorov* theorem [43], for single hidden layer neural networks, if the number of input layer nodes is n, then the number of hidden layer nodes should be at least $2n+1$. In addition, the maximum number of hidden layer nodes is N, which is the number of samples. In this study, the number of input layer nodes is 5 and the number of samples is 20. To take into account the performance and training costs of the networks, this paper adopts the local optimization method to find the number of hidden layer nodes within the range of 11 to 20 by using the testing set so that the ELM model has the best prediction accuracy.

In the optimization process, the input weight matrix (w) and the hidden layer biases (b) of the networks are randomly set. The activation function is separately set as the sigmoid function, sine function and hardlim function. Then, the training set is used to train the networks with different numbers of hidden layer nodes, and the testing set is predicted to output the root mean square error (RMSE) of the predicted results. The results are presented in Figure 5. As seen from Figure 5, when the activation function is the sigmoid function and the number of hidden layer nodes is 13, the RMSE is the smallest, meaning that the ELM model has the best prediction accuracy. Therefore, the activation function is set as the sigmoid function and the number of hidden layer nodes is set to 13.

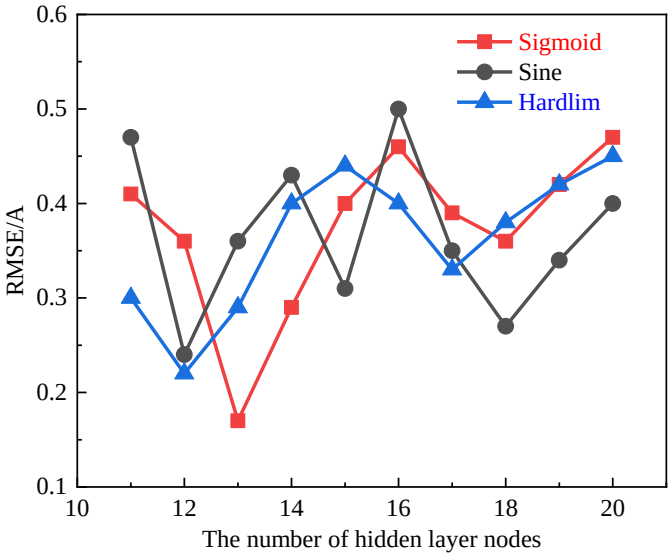

Figure 5. Testing results of the hidden layer nodes and the activation function

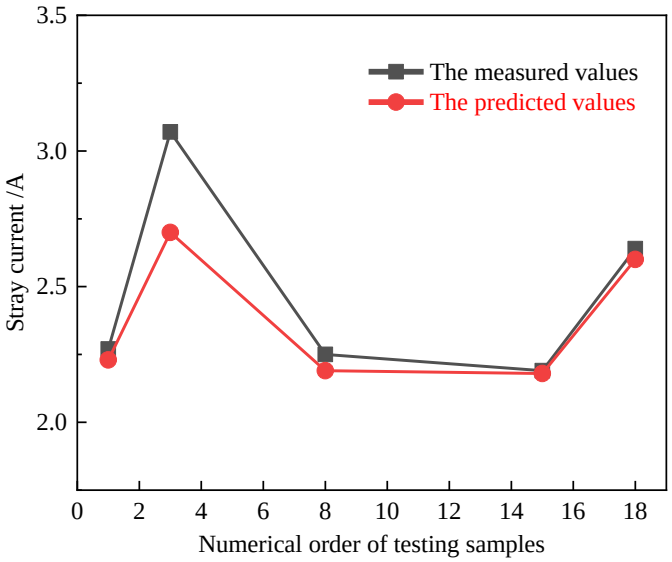

Figure 6. Predicted results of the ELM model

The testing set is input into the trained ELM model to predict the stray current of buried gas pipelines, and the predicted values are shown in Figure 6. In Figure 6, the stray current values predicted by the ELM model are close to the measured values. After calculation, the coefficient of determination of the predicted values of ELM model is 0.922, which is greater than 0.9, indicating that the prediction accuracy and generalization performance of the ELM model are better than those of the MLR model and MNLR model [44-45].

3.5. The PCA-ELM model

Taking the three principal components presented in Table 7 as the input parameters of the ELM model and the stray current of buried gas pipelines as the output parameter, the PCA-ELM neural network model is established using MATLAB (version R2018). The activation function and the number of hidden layer nodes also have strong influences on the prediction accuracy of the PCA-ELM model. Referring to Section 4.4, the optimal number of hidden layer nodes ranges from 7 to 20. Using the same method as described in Section 4.4, the testing results are calculated and presented in Figure 7. The RMSE is the lowest in the case that the activation function is the sine function and the number of hidden layer nodes is 18. Therefore, the activation function of PCA-ELM model is set as the sine function and the number of hidden layer nodes is set to 18.

The predicted values when using the trained PCA-ELM model to predict the stray current of buried gas pipelines are presented in Figure 8. In Figure 8, the stray current values predicted by the PCA-ELM model are very close to the measured values. After calculation, the coefficient of determination of the predicted values of the PCA-ELM model is 0.976, which is close to 1, indicating that the PCA-ELM model has a high prediction accuracy and maintains a good generalization performance for prediction after principal component analysis [46].

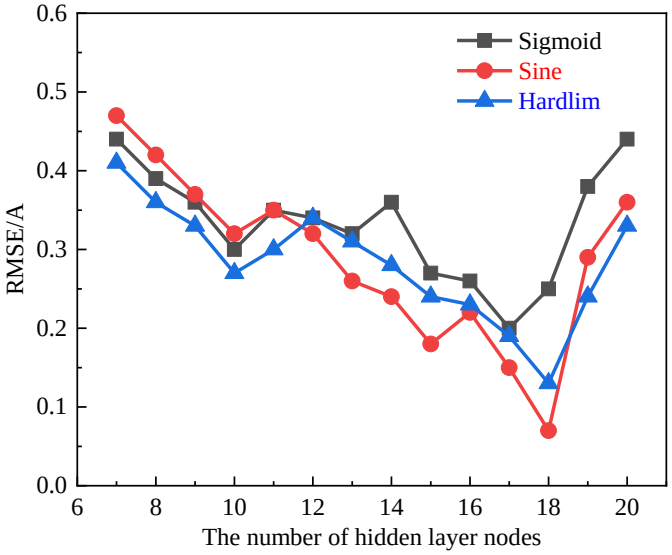

Figure 7. Testing results of the hidden layer nodes and the activation function

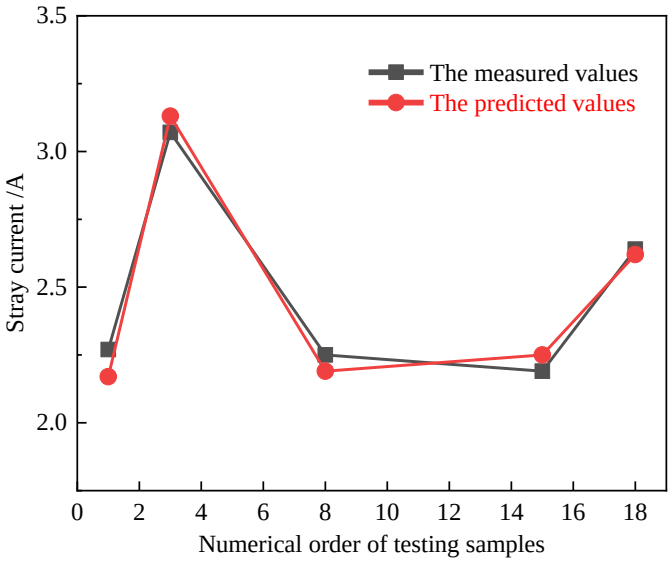

Figure 8. Predicted results of the PCA-ELM model

3.6. Comparison of model prediction results

The MLR model, MNLR model, ELM model and PCA-ELM model are applied to predict the stray current of the testing set, and the predicted values are obtained as shown in Figure 9. Compared with the other models, overall, the predicted values of the PCA-ELM model are the closest to the measured values. After calculation, the coefficients of determination of the predicted values of the MLR model, MNLR model, ELM model and PCA-ELM model are 0.091, 0.368, 0.922 and 0.976, respectively. The higher the coefficient of determination is, the better the prediction accuracy of the model will be. Thus, the preliminary analysis shows that the prediction accuracy of the PCA-ELM model is the best, followed by the ELM model and MNLR model, and the MLR model has a poor

prediction accuracy. Furthermore, the neural network models provide a higher prediction accuracy and better generalization capability than the traditional multiple regression models.

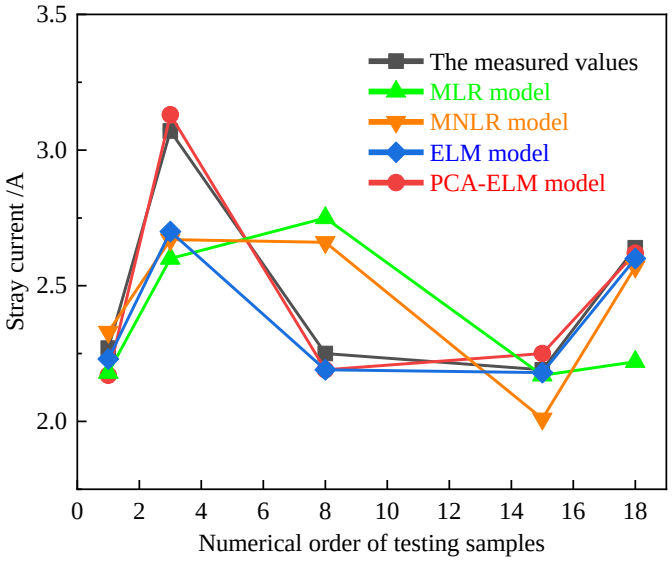

Figure 9. Predicted values of the four models

To further compare the prediction accuracy of the different models, the MAE, MAPE and RMSE are calculated. The results are presented in Table 8. Among the four models, the MAE, MAPE and RMSE of the PCA-ELM model, which are 0.06 A, 2.50% and 0.07 A, respectively, are the smallest. After calculation, these errors are reduced by 80.00%, 78.56% and 80.56%, respectively, compared with the MLR model; decreased by 72.73%, 72.07% and 74.07%, respectively, compared with the MNLR model; and decreased by 40.00%, 32.25% and 58.82%, respectively, compared with the ELM model. Meanwhile, the MAE, MAPE and RMSE of the ELM model are all the second smallest, and those of the MLR model are all the largest.

Table 8. Comparison of the prediction accuracies

The indicators of prediction accuracy	The MLR model	The MNLR model	The ELM model	The PCA-ELM model
MAE/A	0.30	0.22	0.10	0.06
MAPE/%	11.66	8.95	3.69	2.50
RMSE/A	0.36	0.27	0.17	0.07

Therefore, the PCA-ELM model provides the best prediction performance, followed by the ELM model and MNLR model, and the MLR model has a poor prediction accuracy. Cao [10] used the

BP neural network (BPNN) model to predict the stray current density of a buried pipeline, and the MAPE of the BPNN model was calculated to be 5.25%, which was larger than those of the ELM model and PCA-ELM model. Thus, the proposed ELM model and PCA-ELM model both have a higher prediction accuracy than the BPNN model. After similar comparison, the proposed PCA-ELM model has a higher prediction accuracy than the particle swarm optimization neural network (PSO-NN) model to predict the stray current density [11].

4. CONCLUSIONS

In this paper, the multiple linear regression model, multiple nonlinear regression model, extreme learning machine model and extreme learning machine processed by principal component analysis model are utilized to predict the stray current of buried gas pipelines. Based on the analysis of the prediction results, the following conclusions can be drawn.

(1) After analysing the indicators of the prediction accuracy of the four models, the PCA-ELM model is the best, followed by the ELM model and MNLR model, and the MLR model is the worst. Therefore, the PCA-ELM model has more advantages than the other three models in term of the prediction accuracy and provides a reference method for the actual evaluation of the stray current of buried gas pipelines.

(2) Compared with traditional multiple regression models, the neural network models provide a higher prediction accuracy and better generalization performance. Thus, artificial neural networks are potential tools for the prediction of the stray current of buried gas pipelines, guiding us to take necessary measures in a timely manner to reduce stray current corrosion of buried pipelines.

(3) Principal component analysis is combined with an extreme learning machine to forecast the stray current of buried gas pipelines. Using PCA, we obtained three variables that explained more than 80% of the information provided by the original five variables. The predicted results show that PCA not only reduces the data redundancy but also improves the prediction accuracy of the ELM model. In addition, the generalization of the ELM model is also maintained. Thus, PCA can be used as a feasible data processing method for the prediction of stray current or in other prediction studies.

(4) The activation function and the number of hidden layer nodes are sensitive to the prediction accuracy of the ELM model. In this study, the sigmoid function, preferably with 13 hidden layer nodes, is more suitable than the sine function or hardlim function for the ELM model to predict the stray current of buried gas pipelines. However, when the activation function is the sine function and the number of hidden layer nodes is 18, the PCA-ELM model has best prediction accuracy.

ACKNOWLEDGMENTS
We are very grateful for the measuring instruments and guidance provided by Guangzhou Gas Group Co., Ltd.

References

1. L. Bertolini, M. Carsana, and P. Pedeferri, *Corros. Sci.,* 49 (2007) 1056.
2. J. Shi, Y. Zou, J. Ming, and M. Wu, *Corros. Sci.,* 169 (2020) 108610.
3. K. Tang, and S. Wilkinson, *Constr. Build. Mater*., 230 (2020) 117006.
4. Q. Qin, B. Wei, Y. Bai, L. Nan, J. Xu, C. Yu, and C. Sun, *Int. J. Pres. Ves. Pip*., 179 (2020) 104016.
5. K. Zakowski, K. Darowicki, J. Orlikowski, A. Jazdzewska, S. Krakowiak, M. Gruszka, and J. Banas, *Case Studies In Construction Materials*, 4 (2016) 116.
6. H. Bai, *Oil & Gas Science and Technology – Revue d'IFP Energies nouvelles*, 75 (2020) 42.
7. W. Thoe, and J.H.W. Lee, *J. Environ. Eng*., 140 (2014) 472.
8. A.G. More, and S.K. Gupta, *J. Biosci. Bioeng*., 126 (2018) 205.
9. M. Rezaeianzadeh, H. Tabari, A. Arabi Yazdi, S. Isik, and L. Kalin, *Neural Computing And Applications,* 25 (2014) 25.
10. Cao, L. A, Zhu, J. Qing, Zhang, T. Sheng, Hou, and R. Bao, *Anti - Corrosion Methods And Materials*, 57 (2010).
11. C. Wang, W. Li, G. Xin, Y. Wang, and S. Xu, *Complexity,* 2019 (2019) 1.
12. Y. Wan, S. Song, G. Huang, and S. Li, *Neurocomputing*, 260 (2017) 235.
13. Y. Yu, and Z. Sun, *Neurocomputing*, 261 (2017) 50.
14. Y. Peng, W. Kong, and B. Yang, *Neurocomputing*, 266 (2017) 458.
15. G.B. Huang, X. Ding, and H. Zhou, *Neurocomputing*, 74 (2010) 155.
16. Y. Song, S. Zhang, B. He, Q. Sha, Y. Shen, T. Yan, R. Nian, and A. Lendasse, *Neurocomputing*, 277 (2018) 53.
17. L. Wang, X. Li, and Y. Bai, *Energ. Convers. Manage*., 162 (2018) 239.
18. Z. Li, L. Ye, Y. Zhao, X. Song, J. Teng, and J. Jin, *Protection And Control Of Modern Power Systems*, 1 (2016).
19. S. Li, P. Wang, and L. Goel, *Electr. Pow. Syst. Res*., 122 (2015) 96.
20. G. Feng, Z. Qian, and N. Dai, *Neurocomputing*, 82 (2012) 62.
21. Y. Guo, J. Wang, H. Chen, G. Li, J. Liu, C. Xu, R. Huang, and Y. Huang, *Appl. Energ*., 221 (2018) 16.
22. X. Li, H. Xie, R. Wang, Y. Cai, J. Cao, F. Wang, H. Min, and X. Deng, *Neural Comput. Appl*., 27 (2016) 67.
23. Z. Yang, L. Ce, and L. Lian, *Appl. Energ*., 190 (2017) 291.
24. Z. Liu, J. Shao, W. Xu, H. Chen, and Y. Zhang, *Nat. Hazards*, 73 (2014) 787.
25. S. Li, L. Goel, and P. Wang, *Appl. Energ*., 170 (2016) 22.
26. G.B. Huang, H. Zhou, X. Ding, and R. Zhang, *IEEE Transactions on Systems Man & Cybernetics Part B*, 42 (2012) 513.
27. M. Rezghi, and A. Obulkasim, *Expert Syst. Appl*., 41 (2014) 7797.
28. H. Cardot, and D. Degras, *International Statal Review*, 86 (2017) 29.
29. Dobriban and Edgar, *Ann. Stat*., 45 (2016).
30. X.S. Wang, X.J. Ding, and Y.L. Xie, *Advanced Materials Research*, 168-170 (2010) 217.
31. G.B. Huang, L. Chen, and C.K. Siew, *IEEE Transactions On Neural Networks*, 17 (2006) 879.
32. Y.H. Zhang, H. Wang, Z.J. Hu, M.L. Zhang, X.L. Gong, and C.X. Zhang, *Advanced Materials Research,* 608-609 (2012) 564.
33. R. Ahila, V. Sadasivam, and K. Manimala, *Appl. Soft Comput*., 32 (2015) 23.
34. X. Luo, X. Chang, and X. Ban, *Neurocomputing*, 174 (2016) 179.
35. D. Broadhurst, R. Goodacre, A. Jones, J.J. Rowland, and D.B. Kell, *Anal. Chim. Acta*, 348 (1997) 71.
36. I.M.M. Ghani, and S. Ahmad, *Procedia - Social And Behavioral Sciences*, 8 (2010) 549.
37. A.G. Asuero, A. Sayago, and A.G. González, *Crit. Rev, Anal. Chem.,* 36 (2007) 41.
38. Y. Yamamoto, Y. Takahashi, E. Suzuki, N. Mishima, K. Inoue, K. Itoh, Y. Kagawa, and Y. Inoue,

Epilepsy Res., 101 (2012) 202.
39. R.A. Jeffree, S.J. Markich, and A.D. Tucker, *Sci. Total Environ*., 336 (2005) 71.
40. A. Bianchini, and P. Bandini, *Comput-Aided Civ. Inf*., 25 (2010) 39.
41. J. Fan, X. Wang, F. Zhang, X. Ma, and L. Wu, *J. Cleam. Prod.,*., 248 (2020) 119264.
42. X. Bian, S. Li, M. Fan, Y. Guo, and J. Wang, *Anal Methods-Uk*, 8 (2016) 4674.
43. P. Thomas, and M.C. Suhner, *Neural Process. Lett*., 42 (2015) 437.
44. E. Mohammadian, S. Motamedi, S. Shamshirband, R. Hashim, R. Junin, C. Roy, and A. Azdarpour, Environ. Earth Sci., 75 (2016).
45. S.S. Abdullah, M.A. Malek, N.S. Abdullah, O. Kisi, and K.S. Yap, *J. Hydrol.*, 527 (2015) 184.
46. A. Castaño, F. Fernández-Navarro, and C. Hervás-Martínez, *Neural Proccess. Lett*., 37 (2013) 377.