JME

Journal of Mining & Environment,
Vol.7, No.1, 2016, 97-107.
DOI: 10.22044/jme.2016.500

Estimation of Cadmium and Uranium in a stream sediment from Eshtehard region in Iran using an Artificial Neural Network

F. Razavi Rad[1*], F. Mohammad Torab[1] and A.A. Abdollahzadeh[2]

1. Faculty of Mining and Metallurgical Engineering , Yazd University, Yazd, Iran.
2. Department of Computer Engineering, Islamic Azad University, Science and Research Branch, Tehran, Iran.

Received 25 April 2015; received in revised form 28 September 2015; accepted 28 September 2015
**Corresponding author: frazavi@stu.yazd.ac.ir (F. Razavi Rad).*

Abstract
Considering the importance of Cd and U as pollutants of the environment, this study aims to predict the concentrations of these elements in a stream sediment from the Eshtehard region in Iran by means of a developed artificial neural network (ANN) model. The forward selection (FS) method is used to select the input variables and develop hybrid models by ANN. From 45 input candidates, 13 and 14 variables are selected using the FS method for Cadmium and Uranium, respectively. Considering the correlation coefficient (R^2) values, both the ANN and FS-ANN models are acceptable for estimation of the Cd and U concentrations. However, the FS-ANN model is superior because the R^2 values for estimation of Cd and U by the FS-AAN model is higher than those for estimation of these elements by the ANN model. It is also shown that the FS-ANN model is preferred in estimating the Cd and U population due to reduction in the calculation time as a consequence of having less input variables.

Keywords: Artificial Neural Network, Uranium, Cadmium, Forward Selection, Environmental Pollution.

1. Introduction

In the recent years, artificial intelligence (AI) has replaced the traditional scientific methods. Artificial neural network (ANN) is one of the most popular AI ways that uses the mathematical models of the human brain as a system. Neural networks are usually trained with the training data. They can discover new connections, new functions, and new patterns, and have been widely used due to the above characteristics. Nowadays, estimation of the environmental pollutants such as toxic elements (for example, cadmium and uranium) is an important topic in environmental science because it is directly related to the human health. Thus the need for accurate models for their estimation is felt.

In the recent years, ANNs have become extremely popular for estimation and forecasting in a number of areas including finance, power generation, medicine, water resources, and environmental science [1]. The AI-based methods have been proposed as alternatives to the traditional statistical ones in many scientific disciplines. The literature demonstrates that the AI models such as the ANN and neuro-fuzzy ones are successfully used for air pollution modeling [2, 3] and forecasting non-linear phenomena [4, 5]. Carnevale et al. (2009) [6] have presented application of the neural network and neuro-fuzzy models to estimate the non-linear source-receptor relationships between the precursor emissions and pollutant concentrations in Northern Italy.

Input selection is a crucial step in an ANN implementation. This technique is not engineered to eliminate the superfluous inputs. In the case of a high number of input variables, the irrelevant, redundant, and noisy variables might be included in the data set, simultaneously; meaningful variables could be hidden [7, 8]. Therefore, reducing input variables is recommended. There are different methods for reducing the number of input variables such as the forward selection (FS) [9, 10] and gamma test (GT) techniques [11, 12].

In this study, the FS technique was applied in order to build hybrid models with ANN (FS-Cd,

FS-U), and then they were compared with ANN fed with all the input data.

2. Materials and method

2.1. Case study and data

The studied area is a region in the Eshtehard city in the Alborz province, center of Iran, located between the longitudes 50º 00' and 50º 30' E and the latitudes 35º 30' and 36º 00' N, with an area of about 800 km^2 (Figure 1). It is located 62 km from the town of Karaj and 105 km from the city of Tehran. Eshtehard is a relatively desert region, and is located in a semi-arid climate. It neighbors the mountainous cities Halghedare and Nazarabad from the north, and to the mountainous towns Ghezlbash and Malard from the south. It is limited to the Shoor River and the Karaj town from the east, and Buin-zahra town from the west. Currently, it has a population of about 25000.

The sample preparation step was begun with the grinding process. The samples were disaggregated and sieved to <0.18 mm, and then ground to a fine powder (≈0.074 mm).

After sample preparations, the samples were analyzed in the Geosciences Development. They were analyzed for 44 elements. For the high Au values, the analysis method was chosen to be atomic absorption spectrometry, and for the low Au values, emission spectrography was used. The analysis method for Sn was X-ray fluorescence, and for the other elements, it was the ICP-OES method.

2.2. Sampling and chemical analysis

The geochemical samplings were carried out from the stream bed. The samples were air-dried and ground to pass through a 0.18 mm sieve mesh. The number of samples was 357. The sample weights were, on average, about 300 g. In the wet sampling environment, the samples not taken into the sieve.

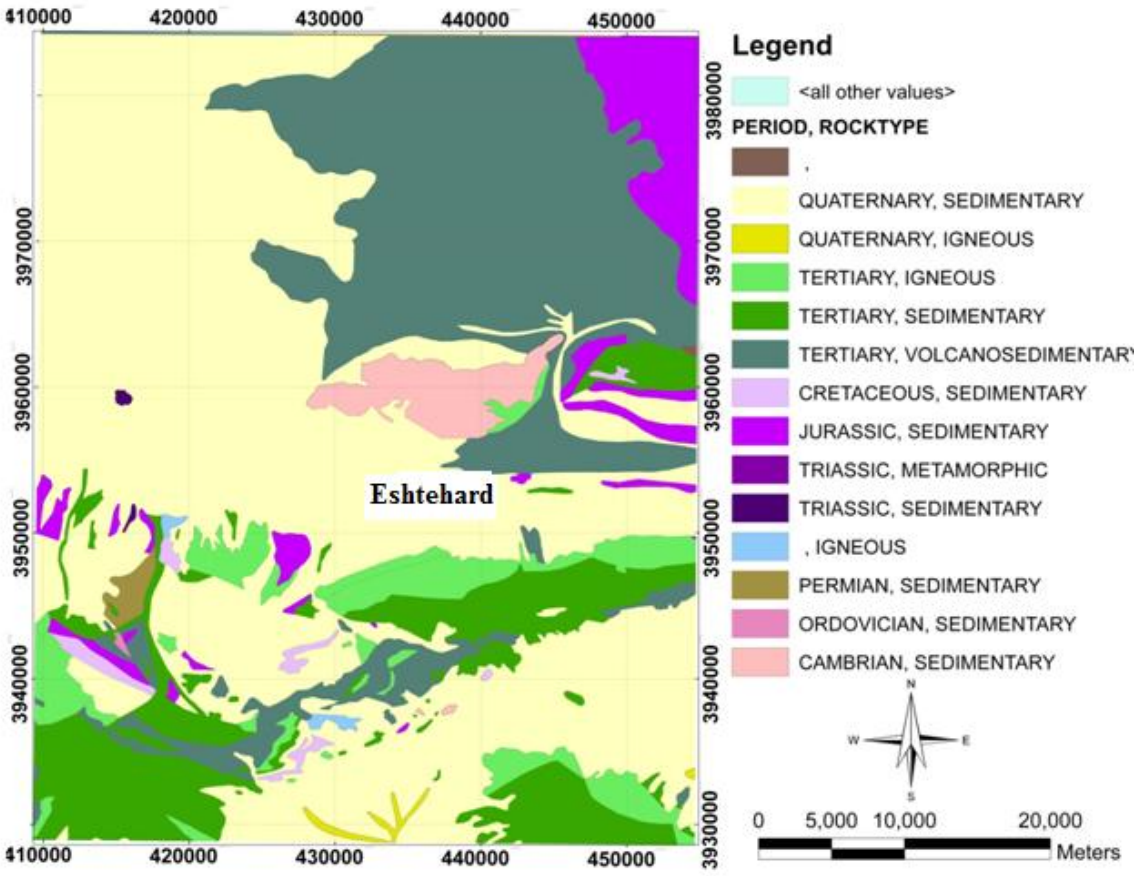


Figure 1. 1:100,000 geological sheet of sampling sites in city of Eshtehard, Alborz Province, Iran.

2.3. Artificial neural networks (ANNs)

The ANN technology was first offered by McCulloch and Pitts in 1943 [13]. Despite the use of a simple structure of this model, its speed and its computing power was highly regarded. ANNs are calculating models that are capable of determining the relationship between the inputs

and outputs of a physical system, although complex and non-linear, with a network of nodes that are interconnected. The important factor that defines ANNs is their architecture. An ANN is a proper mathematical structure that has an interconnected assembly of simple processing elements or nodes. An ANN customary architecture is composed of three layers. Many theoretical and experimental works have shown that a single hidden layer is sufficient for ANNs to approximate any complex non-linear function [14-16]. A major reason for this fact is that the intermediate cells do not directly connect to the output cells. Hence, they would have very small changes in their weight, and learn very slowly [17]. Details of mastering the art of ANN model have been published elsewhere [17, 18]. In this study, a model based on a feed forward neural network with a single hidden layer was used. The back-propagation algorithm was used to train the network. Also the chosen activation functions were sigmoid and linear in the hidden and output layers, respectively.

2.4. Forward selection

When the number of candidate covariates (N) is small, one can choose an estimation model by computing a reasonable criterion (e.g. RMSE, SSE, FPE or cross-validation error) for all the possible subsets of the estimators. However, as N increases, the computational burden of this approach increases very quickly. This is one of the main reasons why step-by-step algorithms like FS are popular. FS has been successfully used by many researchers in order to build robust estimation models [19-21, 10]. In this approach, which is based upon the linear regression model, the first step is ordering the explanatory variables according to their correlation with the dependent variables (from the most to the least correlated variable). Then the explanatory variable, which is best correlated with the dependent variable, is selected as the first input. All the remaining variables are then added one by one as the second input according to their correlation with the output, and the variable that most significantly increases the correlation coefficient (R^2) is selected as the second input. This step is repeated for N-1 times to evaluate the effect of each variable on the model output. Finally, among the N obtained subsets obtained, the subset with optimum R^2 is selected as the model input subset. The optimum R^2 is integral to a set of variables after which, adding a new variable does not significantly increase R^2 [9].

3. Results and discussion

3.1. Input selection

3.1.1. Forward selection

In this study, the FS method was used as a linear input selection technique in order to select the best subset out of 45 input candidates. In other words, a linear model was developed using the best correlated subset of inputs. First, the correlation between each input variable and the desired output was evaluated. Secondly, the variable with the highest correlation, i.e. the concentration of Ag for Cd ($R^2 = 0.49$) and the concentration of V for U ($R^2 = 0.774$) selected as the first and the most important input. Then the remaining candidates were implemented into the model one by one, and the new variable which provided the best modeling result was selected as the new input, and added to the previously selected input. For evaluation of the modeling goodness, the correlation coefficient (R^2) value was used. This step was repeated for several times until adding a new variable to the inputs did not significantly improve the model output. In other words, if increase in R^2 was more than 5%, the new variable was selected. Finally, the input variables with most significant effects on the output were selected, and the other variables were removed. The results obtained for FS were shown, where 13 candidates for Cd (Ag, Cr, Zr, Bi, X, Fe, Ba, As, Ni, y (location), S, P, and Y) and 14 candidates for U (V, Cs, Hg, Bi, Ag, W, Sc, Zr, Mg, Ca, Mn, Mo, Cu, and Ba), according to their importance, were selected as the input variables. The results obtained for the correlation between the estimated data and the observed data for Cd and U are shown in Figures 6b to 9b.

3.2. ANN model development

In this study, the activation function in the hidden layer was a tansigmoid function, and the output value for this function was bounded between -1 and 1; therefore, the input and output data were mapped to [-1, 1]. For evaluating the effect of input selection on the ANN model operation, two models were developed. First, the ANN model was developed using all the input variables, i.e. 45 inputs (ANN model). Secondly, the input variables resulting from the FS method were considered as the ANN inputs, i.e. 13 inputs for cadmium and 14 inputs for uranium (FS-ANN model). To improve the generalization of these models, the stop training algorithm (STA) was used [22]. For implementing STA in practice, the available data was divided into three parts: calibration sets (consisting of the training and

validating set) and testing sets. The results obtained for the correlation between the estimated values and the observed values for Cd and U are shown in Figures 2b to 5b.

Tables 1 and 2 show the results obtained for the calibration and testing of the models with the best structures. According to these tables, although the accuracies of all models are relatively similar, the FS-ANN models are superior because they not only have better accuracy but also have less number of inputs. Among these two models, the FS-ANN model was selected as the best model because of having the least number of inputs.

Table 1. Results of calibrating and testing ANN and FS-ANN models for Cd.

Model	Number of input variables	Calibrating R^2	Testing R^2
ANN	45	0.83	0.75
FS-ANN	13	0.85	0.83

Table 2. Results of calibrating and testing ANN and FS-ANN models for U.

Model	Number of input variables	Calibrating R^2	Testing R^2
ANN	45	0.93	0.90
FS-ANN	14	0.94	0.92

For calculating the values for Cd and U in the Eshtehard region, a numerical code was developed under the MATLAB software. Thus the ANN and FS-ANN models were generated from the data obtained in the region. The results obtained for these models showed that the estimated Cd and U values during the calibrating and testing steps were quantified by estimating the confidence intervals of the simulation results. Plots for the estimates of Cd and U values for the ANN model during the calibration step are shown in Figures 2 and 4, respectively. Also plots of the testing steps are shown in Figure 3 and 5 for Cd and U, respectively. The results obtained for the FS-ANN model during the calibration and testing steps are shown in Figures 6, 7, 8, and 9 for Cd and U (calibrating and testing steps), respectively. According to the results obtained for the ANN and FS-ANN models (Figures 2, 4, 6 and 8), it is obvious that during the calibration stage, both models consistently predicted the trend of decrease and increase in the Cd and U concentrations. A similar trend was also found at the testing step (Figures. 3, 5, 7, and 9). Considering the correlation coefficient (R^2) values, both models (ANN and FS-ANN) were acceptable in estimating the Cd and U concentrations, although FS-ANN was superior. Since the R^2 values for Cd and U estimations in the FS-AAN model were higher than the R^2 values for Cd and U estimations in the ANN model (Tables 1 and 2), in this work, the FS-ANN model was selected as the best estimator for estimation of the Cd and U concentrations.

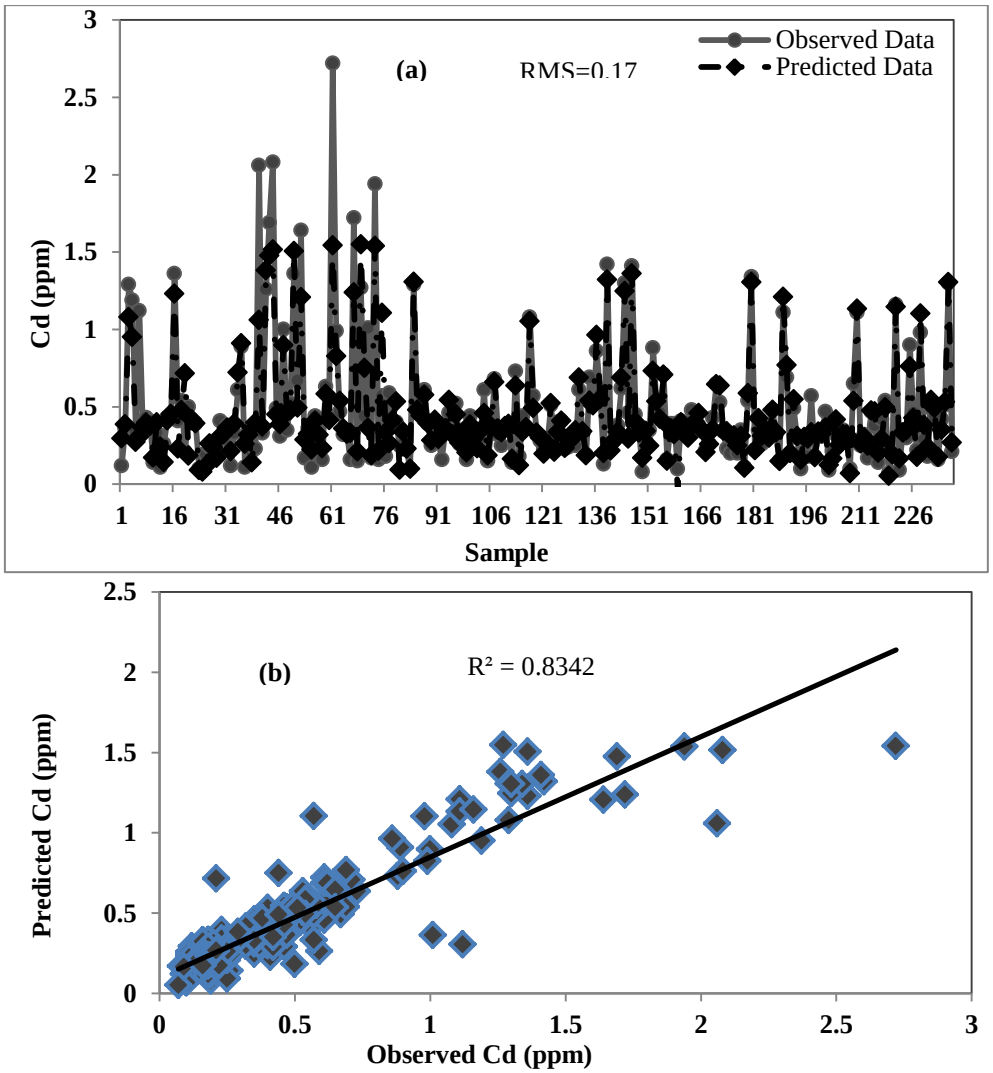
4. Conclusions

Considering the importance of Cd and U concentrations as the pollutants of the environment in the Eshtehard region, located in Iran, this research work aimed to develop proper estimation models using the ANN and FS-ANN models. Since the input selection is a significant step in modeling, the FS method was used, and four models were developed. The goodness of each model was evaluated using the R^2 value. Finally, FS-ANN, as a superior model, was carried out. The input selection improved the estimation capability of the ANN model. It reduced not only the output error but also the calculation time due to having less input variables. The number of selected input variables using FS was 13 for Cd and 14 for U.

Considering the R^2 values, FS-ANN was found to be a superior model, and thus was preferred to ANN (refer to Tables 1 and 2).

Acknowledgments

The authors are grateful to the Geological Survey of Iran. We wish to express our thanks to them for permitting us for the geochemical analyses.



Figures 2. a) 95% confidence intervals for result of modeling for estimates of Cd concentration during calibrating step using ANN model. b) Linear regression between results of observed Cd and estimated Cd during calibrating step.

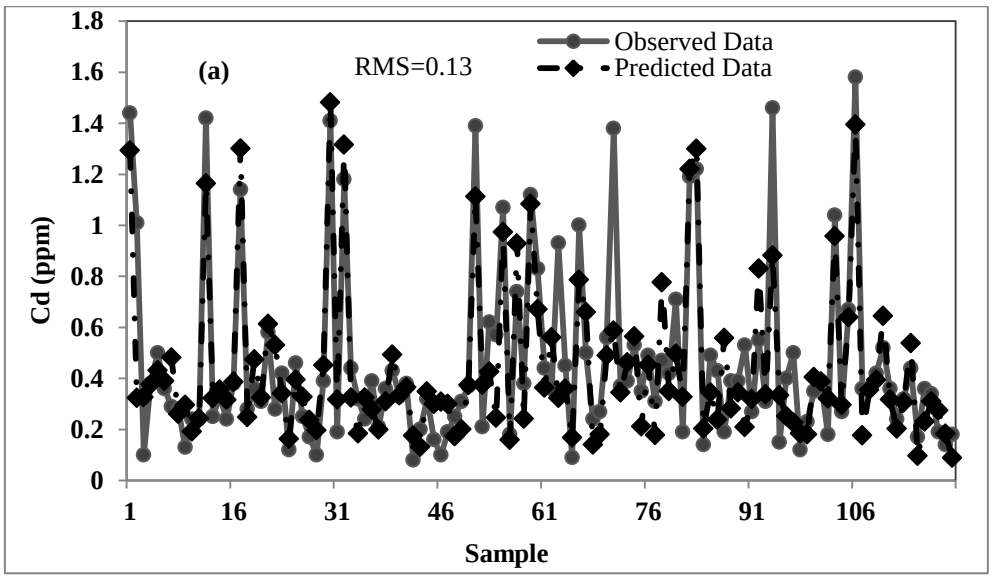


Figure 3. a) 95% confidence intervals for result of modeling for estimates of Cd concentration during testing step using ANN model. b) Linear regression between results of observed Cd and estimated Cd during testing step.

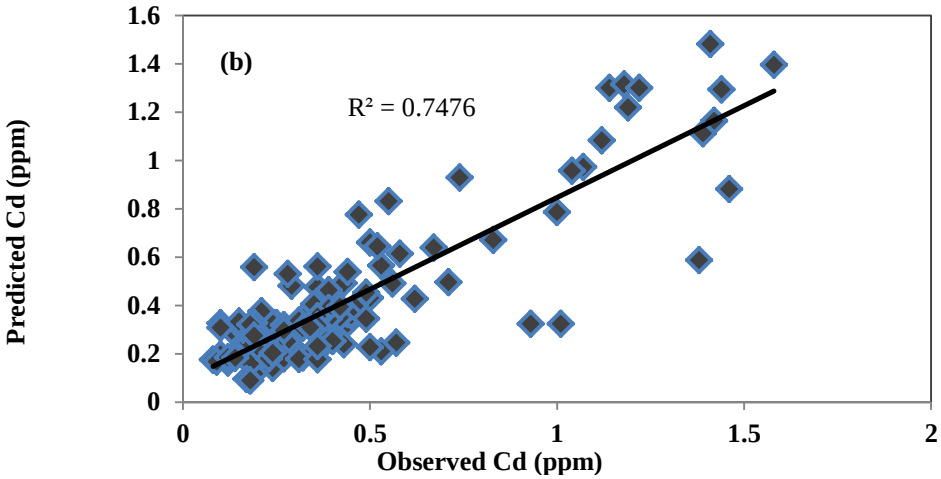
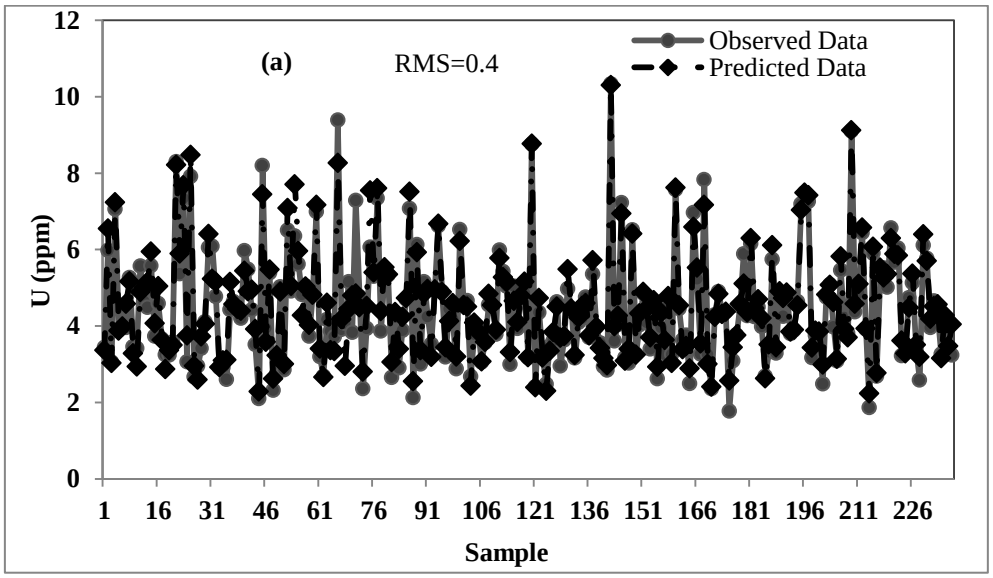


Figure 3. Continued.



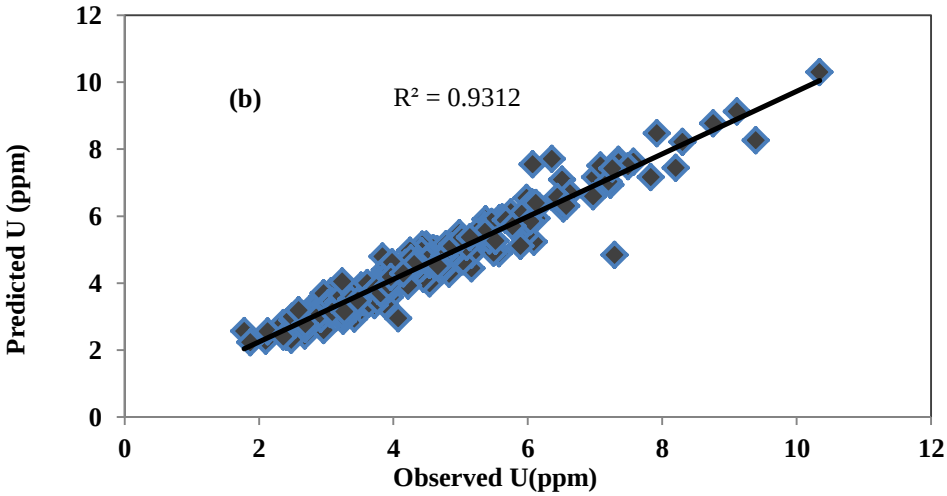


Figure 4. a). 95% confidence intervals for result of modeling for estimates of U concentration during calibrating step using ANN model. b) Linear regression between results of observed U and estimated U during calibrating step.

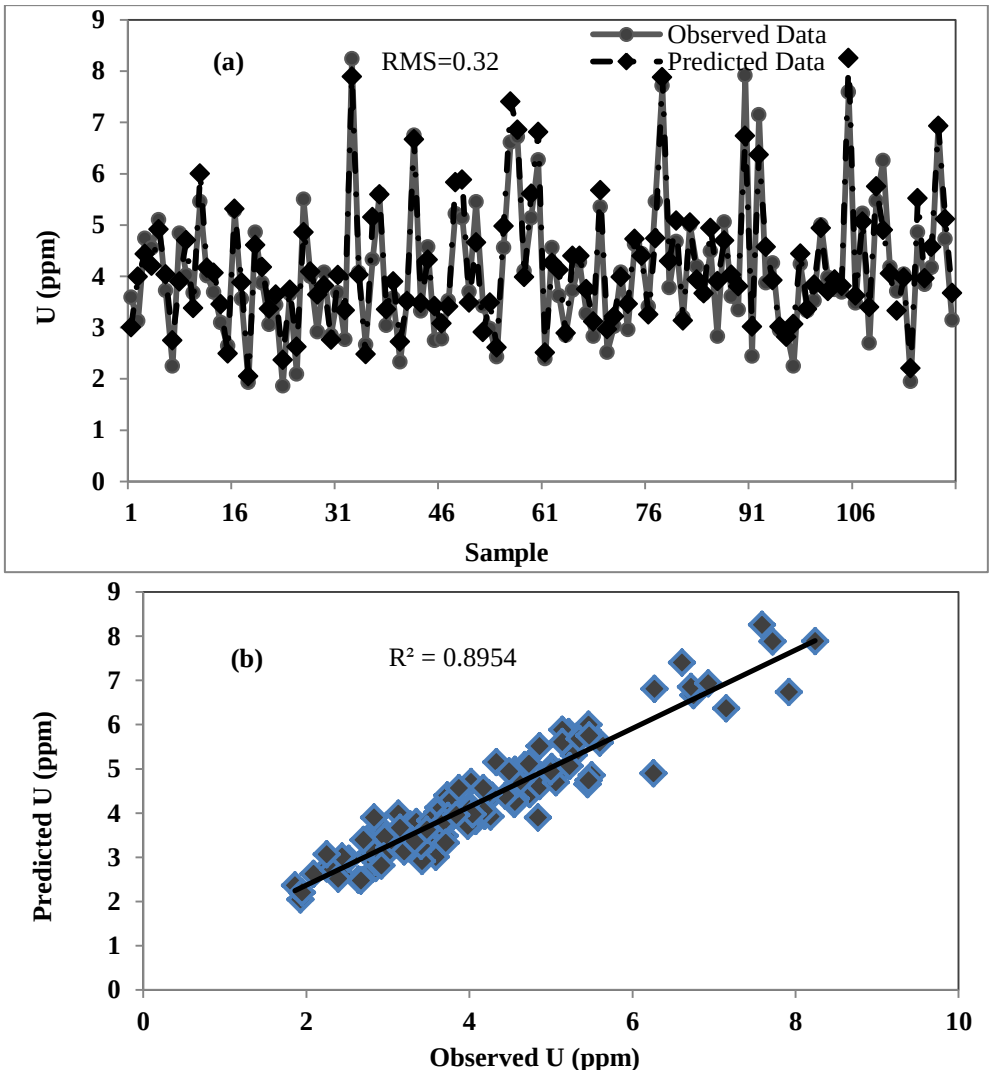


Figure 5. a) 95% confidence intervals for result of modeling for estimates of U concentration during testing step using ANN model. b) Linear regression between results of observed U and estimated U during testing step.

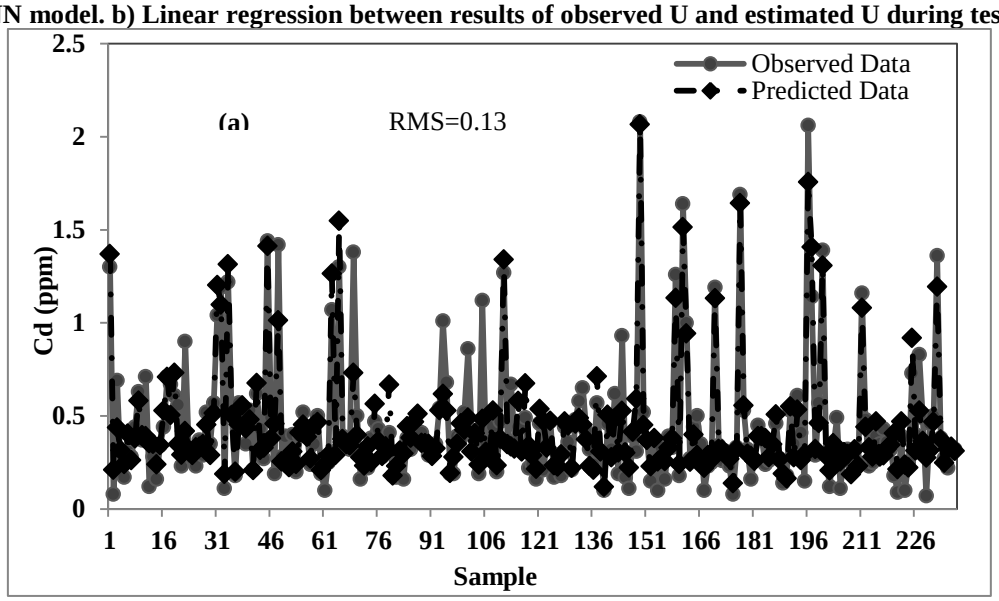


Figure 6. a) 95% confidence intervals for result of modeling for estimates of Cd concentration during calibrating step using FS-ANN model. b) Linear regression between results of observed Cd and estimated Cd during calibrating step.

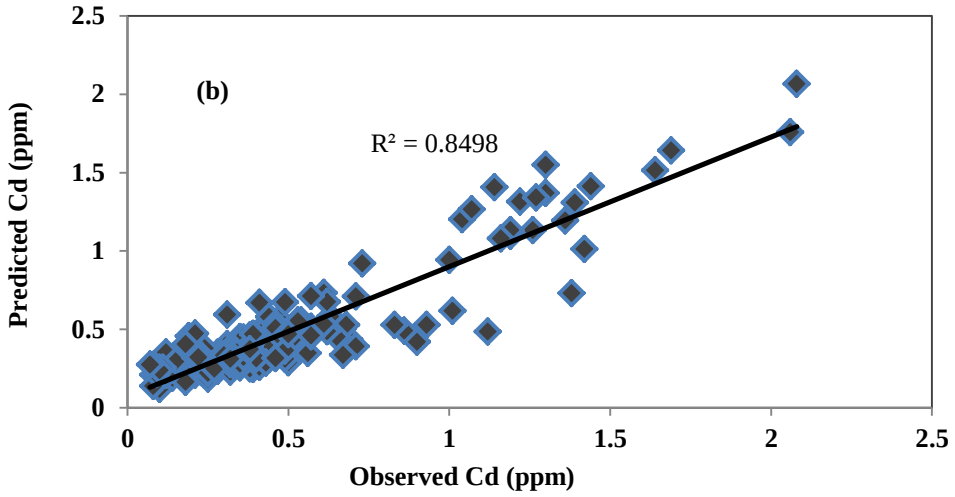


Figure 6. Continued.

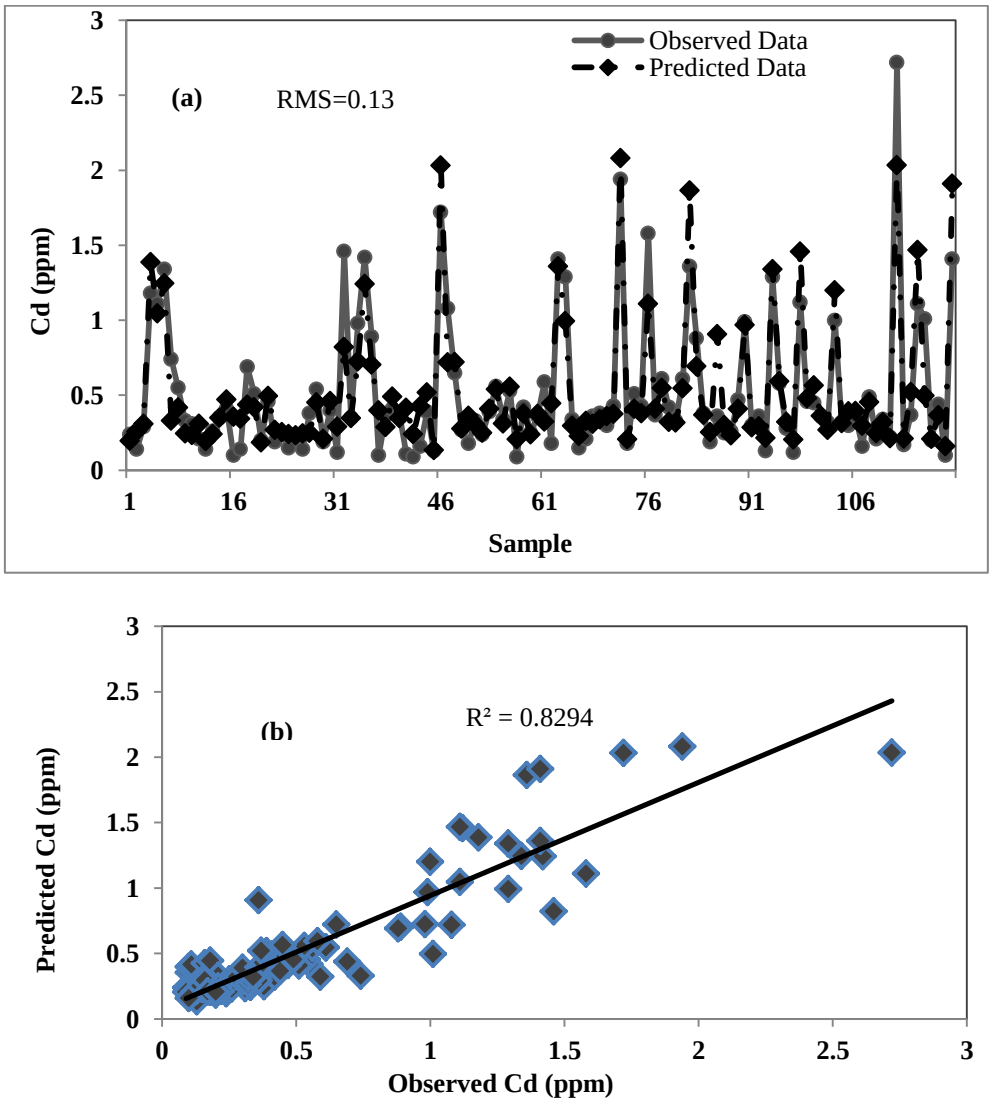


Figure 7. a) 95% confidence intervals for result of modeling for estimates of Cd concentration during testing step using FS-ANN models. b) Linear regression between results of observed Cd and estimated Cd during testing step.

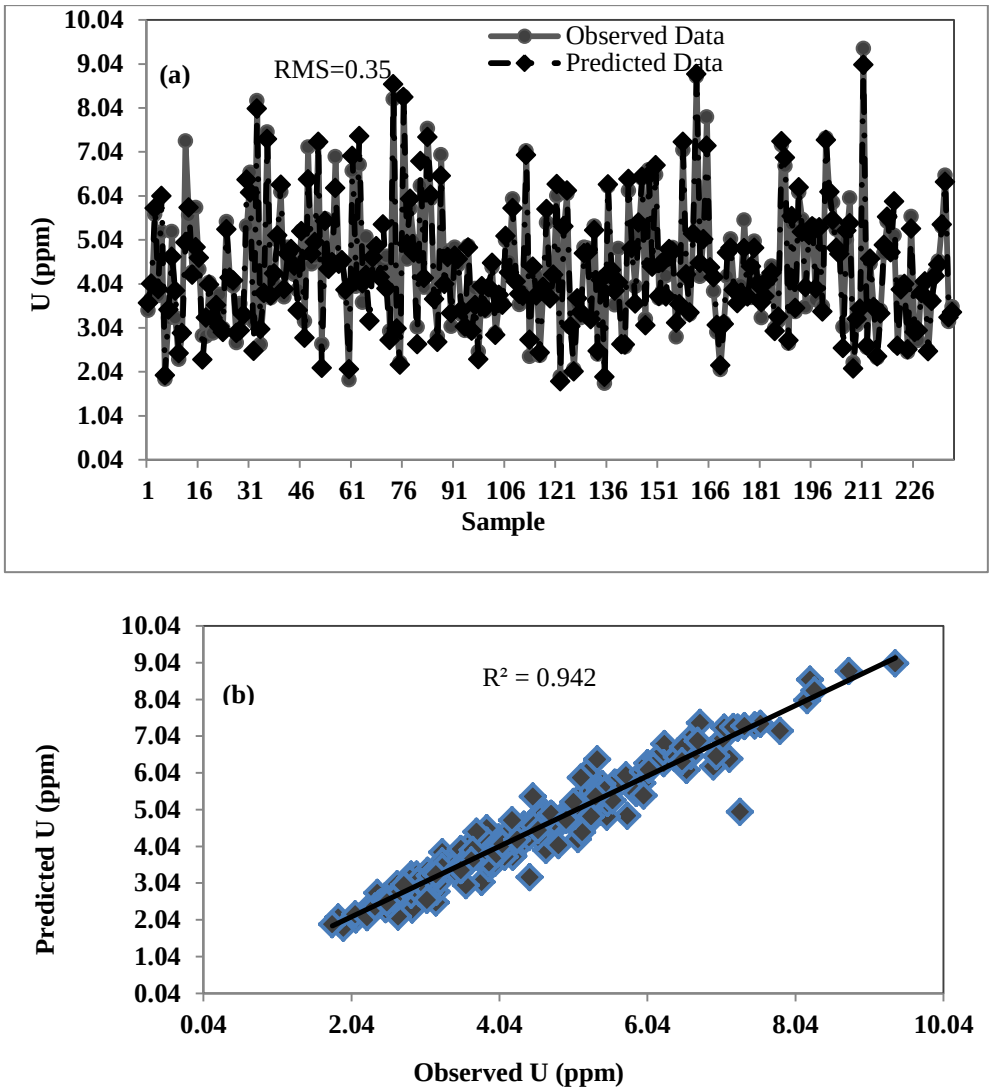


Figure 8. a) 95% confidence intervals for result of modeling for estimates of U concentration during calibrating step using FS-ANN model. b) Linear regression between results of observed U and estimated U during calibrating step.

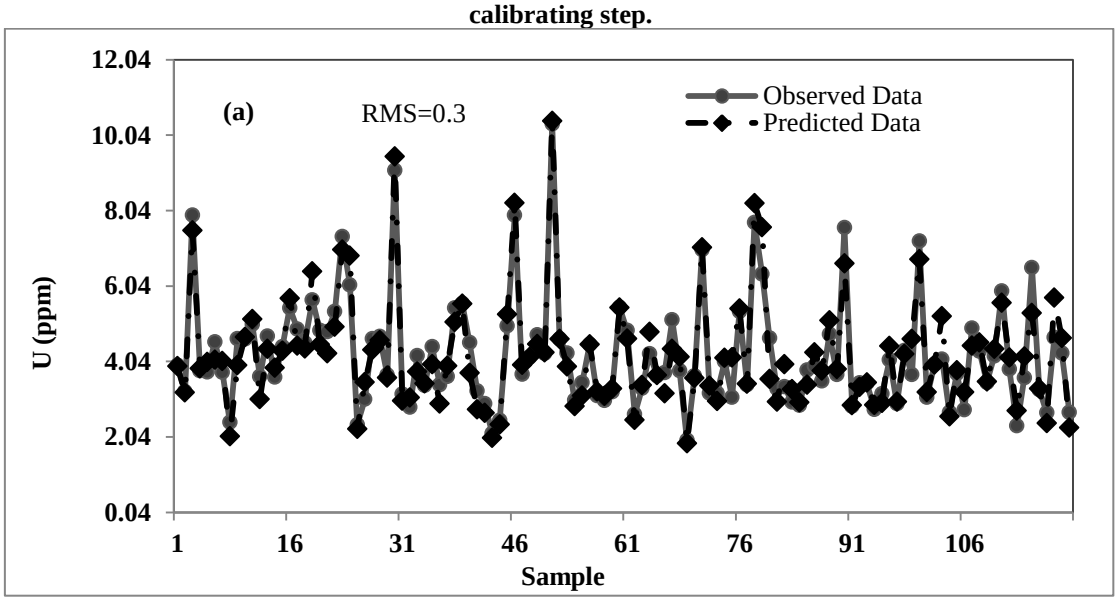


Figure 9. a) 95% confidence intervals for result of modeling for estimates of U concentration during testing step using FS-ANN model. b) Linear regression between results of observed U and estimated U during testing step.

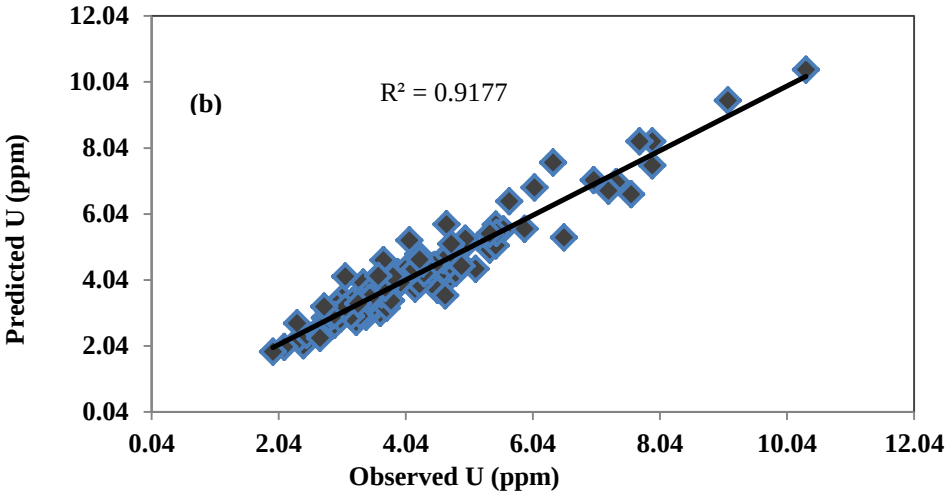


Figure 9. Continued.

References

[1]. Maier Holger, R. and Dandy Graeme, C. (2000). Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications., Environmental Modelling & Software. 15: 101-124.

[2]. Nunnari, G., Dorling, S., Schlink, U., Cawley, G., Foxall, R. and Chatterton, T. (2004). Modelling SO2 concentration at a point with statistical approaches., Environmental Modelling & Software. 19: 887-905.

[3]. Castroa, R., Jorqueraa, J., Perez-Correaa, H., Vesovic, J.R. and Perez-Roaa, V. (2006). Airpollution modeling in an urban area: correlating turbulent diffusion coefficients by means of an artificial neural network approach., Atmospheric Environment. 40:109-125, 2006.

[4]. Trier, P., Reyes, A. and Perez, J. (2000). Prediction of PM 2.5 concentrations several hours in advance using neural networks in Santiago, Chile., Atmospheric Environment. 34: 1189-1196.

[5]. Gautama, A.K., Chelanib, A.B., Jaina, V.K. and Devottab, S. (2008). A new scheme to predict chaotic time series of air pollutant concentrations using artificial neural network and nearest neighbor searching., Atmospheric Environment. 42: 4409-4417.

[6]. Carnevale, C., Finzi, G., Pisoni, E. and Volta, M. (2009). Neuro-fuzzy and neural network systems for air quality control., Atmospheric Environment. 43: 4811-4821.

[7]. Seasholtz, M.B. and Kowalski, B. (1993). The parsimony principle applied to multivariate calibration., Analytica Chimica Acta. 277: 165-177.

[8]. Noori, R., Abdoli, M.A., Ameri, A. and Jalili-Ghazizadeh, M. (2009). Prediction of municipal solid waste generation with combination of support vector machine and principal component analysis: a case study of Mashhad., Environmental Progress & Sustainable Energy. 28: 249-258.

[9]. Chen, S., Billings, S.A. and Luo, W. (1989). Orthogonal least squares methods and their application to nonlinear system identification., International Journal of Control. 50: 1873-1896.

[10]. Wang, X.X., Chen, S., Lowe, D. and Harris, C.J. (2006). Sparse support vector regression based on orthogonal forward selection for the generalised kernel model., Neurocomputing. 70: 462-474 .

[11]. Corcoran, J., Wilson, I. and Ware, J. (2003). Predicting the geo-temporal variation of crime and disorder., International Journal of Forecasting. 19: 623-634.

[12]. Moghaddamnia, A., Ghafari-Gousheh, M., Piri, J., Amini, S. and Han, D. (2009). Evaporation estimation using artificial neural networks and adaptive neuro-fuzzy inference system techniques., Advances in Water Resources. 23: 88-97.

[13]. McCulloch, W.S., and Pitts, W. (1943). A logical calculus of the ideas imminent in nervous activity., B. Math. Biophys. 8: 115-133.

[14]. Cybenko, G. (1989). Approximation by superposition of a sigmoidal function., Mathematics of Control, Signals, and Systems. 2: 303-314.

[15]. Jalili-Ghazizadeh, M. and Noori, R. (2008). Prediction of municipal solid waste generation by use of artificial neural network: a case study of Mashhad., International Journal of Environmental Research. 2: 13-22.

[16]. Noori, R., Abdoli, M.A., Jalili-Ghazizadeh, M. and Samifard, R. (2009). ComparisonofANNand PCA based multivariate linear regression applied to predict the weekly municipal solid waste generation in

Tehran., Iranian Journal of Public Health. 38: 74-84.

[17]. Gallant, S.I. (1993). NeuralNetwork Learning and Expert Systems. Cambridge: MIT Press.

[18]. Haykin, S. (1994). Neural Networks: a Comprehensive Foundation. New Jeresy: Prentice Hal.

[19]. Chen, S., Hong, X., Harris, C.J. and Sharkey, P.M. (2004). Sparse modeling using orthogonal forward regression with PRESS statistic and regularization., IEEE Transactions on Systems, Man, and Cybernetics e Part B. 34: 898-911.

[20]. Eksioglu, B., Demirer, R. and Capar, I. (2005). Subset selection in multiple linear regression: a new mathematical programming approach., Computers & Industrial Engineering. 49: 155-167.

[21]. Khan, J.A., Aelst, S.V. and Zamar, R.H. (2007). Building a robust linear model with forward selection and stepwise procedures., Computational Statistics & Data Analysis. 52: 239-248.

[22]. Coulibaly, P., Ancti, F. and Bobee, B. (2000). Daily reservoir inflow forecasting using artificial neural networks with stopped training approach., Journal of Hydrology. 230: 244-257.

شبیه‌سازی کادمیوم و اورانیوم موجود در رسوبات آبراهه‌ای منطقه اشتهارد با استفاده از روش شبکه عصبی

فاطمه رضوی‌راد[1*]، فرهاد محمد تراب[1] و علی‌اصغر عبداله زاده[2]

۱- دانشکده مهندسی معدن و متالورژی، دانشگاه یزد، ایران

۲- دانشکده مهندسی کامپیوتر، واحد علوم و تحقیقات تهران، دانشگاه آزاد اسلامی، ایران

* نویسنده مسئول مکاتبات: frazavi@stu.yazd.ac.ir

ارسال ۲۰۱۵/۴/۲۵، پذیرش ۲۰۱۵/۹/۲۸

چکیده:

با توجه به اهمیت کادمیوم و اورانیوم به‌عنوان آلوده‌کننده‌های محیط‌زیست، این مطالعه با استفاده از روش شبکه عصبی به پیش‌بینی میزان این عناصر در منطقه‌ی اشتهارد می‌پردازد. روش انتخاب پیشرو به‌منظور انتخاب متغیرهای ورودی مؤثر بر روی پیش‌بینی مقادیر عناصر کادمیوم و اورانیوم و کاهش تعداد کل متغیرها استفاده شد. از تعداد ۴۵ متغیر ورودی اولیه، به ترتیب ۱۳ و ۱۴ متغیر مؤثر بر روی پیش‌بینی مقادیر کادمیوم و اورانیوم توسط روش انتخاب پیشرو انتخاب شدند. با توجه به مقادیر ضریب همبستگی، هر دو مدل (شبکه عصبی و انتخاب پیشرو) برای پیش‌بینی مقادیر کادمیوم و اورانیوم در منطقه مناسب هستند؛ اما روش انتخاب پیشرو مناسب‌تر است، زیرا مقادیر ضریب همبستگی روش انتخاب پیشرو برای پیش‌بینی کادمیوم و اورانیوم بالاتر از روش شبکه عصبی است. هم‌چنین در این مطالعه نشان داده شده است که مدل انتخاب پیشرو در پیش‌بینی مقادیر کادمیوم به دلیل کاهش زمان محاسبات درنتیجه‌ی تعداد متغیرهای ورودی کمتر مناسب‌تر است.

کلمات کلیدی: روش شبکه‌ی عصبی، اورانیوم، کادمیوم، روش انتخاب پیشرو، آلودگی زیست‌محیطی.