Machine Learning Algorithm for Prediction of Heavy Metal Contamination in the Groundwater in the Arak Urban Area

Feridon Ghadimi [1,*]

1- Associated of professor, Department of Mining Engineering, Arak University of Technology, Arak, Iran PO BOX 38181-41167.

* Corresponding Author: ghadimi@arakut.ac.ir

Received: 20 August, 2016 / Accepted: 30 February 2017 / Published online: 05 March 2017

Abstract

This paper attempts to predict heavy metals (Pb, Zn and Cu) in the groundwater from Arak city, using support vector regression model(SVR) by taking major elements (HCO_3, SO_4) in the groundwater from Arak city. 150 data samples and several models were trained and tested using collected data to determine the optimum model in which each model involved two inputs and three outputs. This SVR model fit captures the prime idea of statistical learning theory in order to obtain a good forecasting of the dependence among the major elements in the city of Arak. Finally, on the basis of these numerical calculations using SVR model, from the experimental data, conclusions of this study are exposed. By comparison between the predicted and the measured data it indicates that SVR model has strong potential to estimation of the heavy metals in the groundwater with high degree of accuracy.

Keywords: Groundwater; Support vector regression; Heavy metals; Arak.

1- Introduction

Many factors control the chemical composition of groundwater, which include the composition of precipitation, mineralogy of watersheds and geochemical processes (Andre *et al*., 2005; Singh *et al*., 2014; Zghibi *et al*., 2014). These processes effect on water quality and are responsible for variations in the groundwater composition (Helstrupe *et al*., 2007; Monjerezi *et al*., 2011; Yidana 2010; Anderson *et al*., 2014). Quality of water depends not only on chemical and physical properties of surrounding rocks but also varies as a result of human activity (Monjerezi *et al*., 2011; Matiatos *et al*., 2014; Devic *et al*., 2014; Zapata *et al*., 2014). Hydro-chemical processes, including dissolution, precipitation, weathering together with residence time occurring along flow path, control variation in chemical composition of groundwater (Oinam *et al*., 2012; Wang *et al*., 2013; Srinivasamoorthy *et al*., 2014; Masoud 2014).

Moreover, over the years, the application of artificial neural network (ANN) in different fields of engineering has been developing. An artificial neural network (ANN), usually termed neural network (NN), is a mathematical model or computational model that is based on the structure and functional aspects of biological neural networks. The use of the artificial neural networks (Haykin 1999; Hassan *et al*., 2014) of multilayer perceptron (MLP) type as the model of pollution was exploited frequently in the last years (Aguirre-Basurko *et al*., 2006). In this research work, it is proposed the system focused on the support vector machines (SVM) due to their versatility to tackle complex and highly nonlinear problems with success (Bishop, 2006;

Shawe-Taylor and Cristianini, 2004; Chen, 2015). The SVM networks are built for the prediction of each considered heavy metals (Pb, Zn and Cu) in the groundwater from Arak city. On the other hand, similar to conventional feed-forward (FF) neural networks (NN), the SVM has been used by researchers to solve classification and regression problems (Suárez Sánchez *et al*., 2011a, b). Possessing similar universal approximation ability, SVR can also be used to model nonlinear processes. Compared with the FFNN models, the SVR model has certain advantages. In the first place, training for the SVR gives place to a global optimum. This is due to the fact that SVR is formulated as a convex quadratic optimization problem for which there is a global optimum (Deng *et al*., 2012). On the other hand, the training of (FF) and (NNs) may become trapped at a local minimum. Therefore, mathematically, the SVR model has more attractive properties than the NN model. The second advantage is that the design and training for the SVR model are relatively more straightforward and systematic as compared with those for the NN model. The third advantage is that it is relatively easier to achieve good generalization when using SVR as compared with NNs. Finally, the SVR is a type of model that is optimized so that prediction error and model complexity are simultaneously minimized. To fix ideas the formulation of SVR captures the main finding of statistical learning theory in order to obtain a good generalization so that both training error and model complexity are controlled, by explaining the data with a simple model (Deng *et al*., 2012; Abbasi *et al*., 2013). The objectives of this study were as follow: 1) to explore applications of a support vector machines(SVR) methods in predicting heavy metals in groundwater 2) to develop a model based on support vector machines and evaluate the applicability of the SVR approach to assess and predict heavy metals in groundwater.

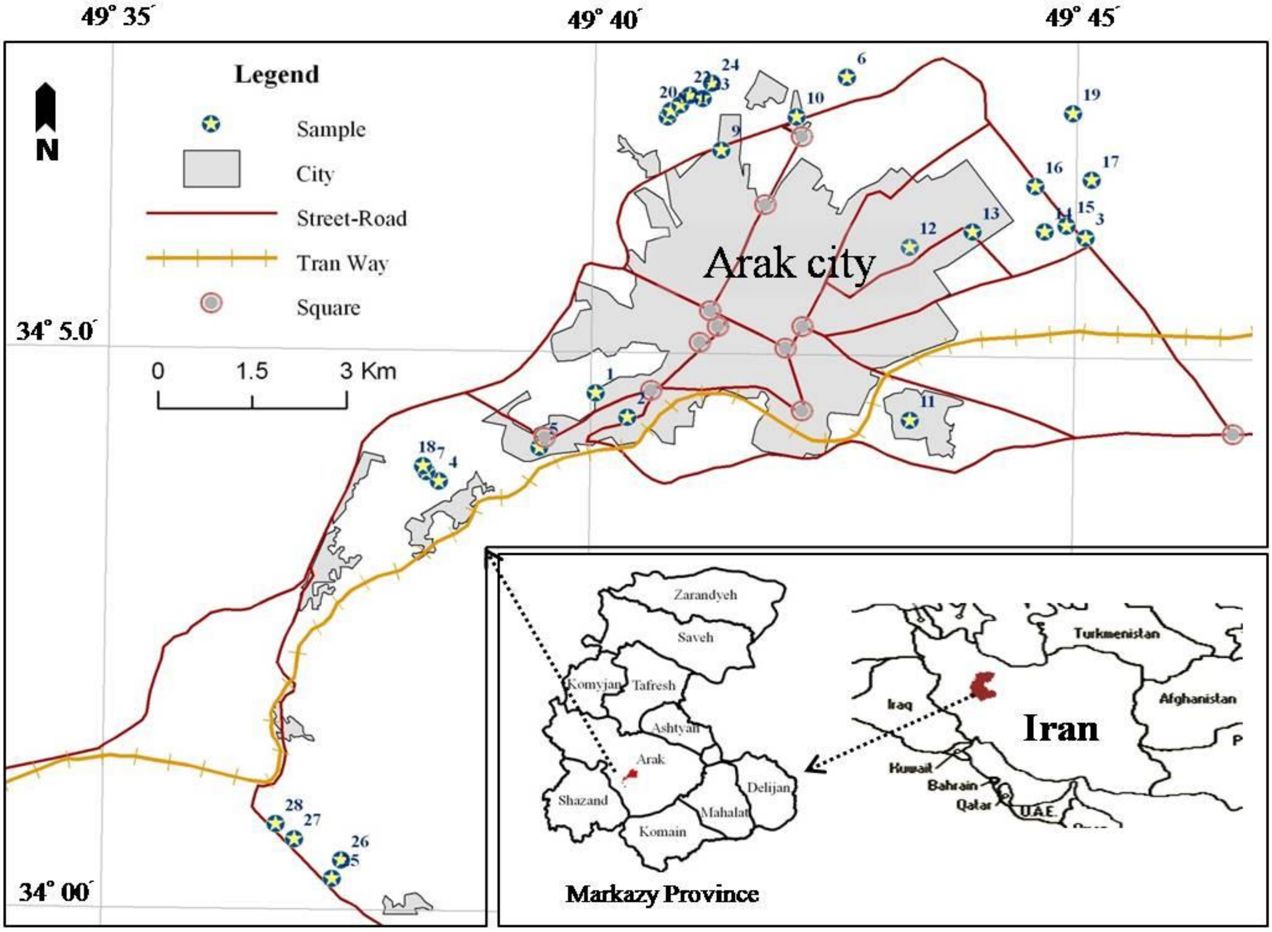

Figure 1) Location map of the study area showing and some of sample locations.

2- Materials and methods

2.1- Area Descriptions

Arak is characterized by a semi-arid climate and an average precipitation and temperature of about 280 mm/year and 11 oC, respectively (Zamani, 1999). Most of its inhabitants are concentrated in town of Arak with more than 600000 inhabitants and work mainly in the industrial plants (Figure1). The study area is situated in the alluvial plain and aquifer is directly fed by a stream of water coming from different reliefs surrounding the depression inter-mountainous of Mighan playa. The plain hosts a large number of water-wells with depths varying from 70 to 150 m. The direction of groundwater flow around Arak plain is from southwest to northeast and toward saline Mighan playa. Arak is one of the regions that its groundwater affected by contamination of industrial origin. The Arak is one of the industrial regions in Iran where the impact of rapid population growth and industrialization on limited natural sources and agricultural lands is progressively high and as a result, the size of uncontaminated areas is being diminished. Due to expanding industrialization and urbanization in Arak and the unrestrained disposal of factory wastes to groundwater, it is thought that heavy metal contents in this region are high. Therefore, monitoring of this change and determination of contamination in the groundwater has gained importance.

2.2- Groundwater sampling

Water samples were collected from shallow wells for urban water supply using standard sampling procedures during sampling campaigns in 2014. The shallow wells were drilled to depths between 70 and 150 m. Total of 150 samples were taken for this study. Samples were collected in 250 ml sterilized polythene bottles. All samples were analyzed for main chemical descriptors using standard methods. Parameters analyzed include major ions of calcium (Ca), magnesium (Mg), potassium (K), sodium (Na), chloride (Cl), sulfate (SO_4) in milligram per liter using ion chromatograph (I.C). Bicarbonate ion concentration in water was determined by titration. Heavy metals were determined by Graphite Furnace Atomic Absorption Spectrophotometer (Perkin–Elmer Analyst 700) using multi element Perkin–Elmer standard solutions. Accuracy of chemical analysis was verified by calculating ion-balance errors where errors were generally within 10%.

2.3- Support vector regression

Let us consider a simple linear regression problem trained on data set $\chi = \{u_i, v_i; i = 1,...,n\}$ with input vectors u_i and linked targets v_i. A function g(u) has to be formulated approximately in order to link up inherited relations between the data sets and thereby it can be used in the later part to infer the output v for a new input data u. Standard SVM regression uses a loss function Lε (v, g(u)) which describes the deviation of the estimated function from the original one. Several types of loss functions can be mined in the literature e.g., linear, quadratic, exponential, Huber's loss function, etc. In the present context the standard Vapnik's – ε insensitive loss function is used which is defined as Eq.(1):

$$L\varepsilon(v, g(u)) = \begin{cases} o & for|v-g(u)\leq\varepsilon| \\ |v-g(u)|-\varepsilon & otherwise \end{cases} \tag{1}$$

Using ε-insensitive loss function, one can find g(u) that can better approximate the actual output vector v and has the at most error tolerance ε from the actual incurred targets v_i for all training data, and concurrently as flat as possible(Fletcher1987). Consider the regression function defined by Eq. (2):

$$g(u) = w\,.u + b \tag{2}$$

where w ∈ χ, χ is the input space; b ∈ R is a bias term and (w. u) is dot product of vectors w and u. flatness in Eq. (2) refers to a smaller value of parameter vector w. By minimizing, the norm $\|w\|^2$ flatness can be ascertained along with model complexity. Thus regression problem can be stated as the following convex optimization problem (Eq. (3)):

$$\min_{w,b,\zeta,\zeta_i^*} \frac{1}{2}\|w\|^2 + c\sum_{i=1}^{N}\left(\zeta_i + \zeta_i^*\right) \qquad (3)$$

$$v_i - (w\, u_i + b) \le \varepsilon + \zeta_i$$

Subject to

$$(w\, u_i + b) - v_i \le \varepsilon + \zeta_i^*$$

$$\zeta_i, \zeta_i^* \ge 0, i = 1.2, ...n$$

where ζ_i and ζ_i^* are slack variables introduced to evaluate the deviation of training samples outside ε-insensitive zone. The trade-off between the flatness of g and the quantity up to which deviations greater than ε are tolerated is depicted by C > 0. C is a positive constant influencing the degree of penalizing loss when a training error occurs. Under fitting and over fitting of training data are avoided by minimization of the regularization term $w^2/2$ along with the training error term $c\sum_{i=1}^{n}\left(\zeta_i - \zeta_i^*\right))$ in Eq. (3). The minimization problem in Eq. (3) represents the primal objective function. Now the problem is dealt by constructing a Lagrange function from the primal objective function by introducing a dual set of variables, $\underline{\alpha}_i$ and $\overline{\alpha}_i$ for the corresponding constraints. Optimality conditions are exploited at the saddle points of a Lagrange function leading to the formulation of the dual optimization problem (Eq. (4)):

$$\max_{\underline{\alpha}_i . \alpha_i} \qquad (4)$$

$$-\frac{1}{2}\sum_{i,j-1}^{n}\underline{\alpha}_j - \overline{\alpha}_i\left(\underline{\alpha}_j - \overline{\alpha}_i\right)\left(u_i . u_j\right) - \varepsilon\sum_{i=1}^{n}\left(\underline{\alpha}_j + \overline{\alpha}_i\right) + \sum_{i=1}^{n} v_i\left(\underline{\alpha}_j - \overline{\alpha}_i\right)$$

Subject to $\sum_{i=1}^{n}\left(\underline{\alpha}_i - \overline{\alpha}_i\right) = 0$ (5)

$$0 \le \underline{\alpha}_i \le c, i = 1, 2, ..., n$$

$$0 \le \overline{\alpha}_i \le c, i = 1, 2, ..., n$$

After determining Lagrange multipliers $\underline{\alpha}_i$ and $\overline{\alpha}_i$ the parameter vectors w and b can be evaluated under Karush–Kuhn–Tucker (KKT) complementarily conditions which are not discussed herein (Fletcher, 1987). Therefore, the prediction is a linear regression function that can be expressed as Eq. (5):

$$g(u) = \sum_{i=1}^{n}\alpha_i - \overline{\alpha}_i \langle u_i\, u\rangle + b \qquad (5)$$

Thus SVM regression expansion is derived; where w is depicted as a linear combination of the training patterns v_i and b can be found using primary constraints. For |g(u)|≥ε Lagrange multipliers may be non-zero for all the samples inside the ε-tube and these remaining coefficients are termed as support vectors.

Now for making SVM regression to deal with non-linear cases; pre-processing of training patterns u_i has to done by mapping the input space χ into some feature space $\Im$ using nonlinear function $\varphi = \chi \rightarrow \Im$ and is then applied to the standard support vector algorithm. Also the dimensionality of $\varphi_{(x)}$ can be very huge, making 'w' hard to represent explicitly in memory, and hard for the quadratic programming optimizer to solve. The theorem Kimeldorf and Wabha shows that:

$w = \sum_{i=1}^{n}\alpha_i . \phi(x_i)$ for some variables. Instead of optimizing 'w' we can directly optimize α_i. There by the decision function is obtained (kernel trick1). The theorem is exploited to examine the sensitivity properties of ε-insensitive SVR and introduce the concept of approximate degrees of freedom (Fletcher, 1987). The degrees of freedom play a vital role in the assessment of the optimism i.e., the difference between the expected in sample error and the expected empirical risk. Let u_i be mapped into the feature space by nonlinear function $\varphi(u)$ and hence the decision function is given by Eq. (6):

$$g(w,b) = w.\phi(u) + b \qquad (6)$$

This nonlinear regression problem can be expressed as the following optimization problem. Figure 2 depicts the concept of non-linear SV regression corresponding to Eq. (7).

$$\min_{w,b,\zeta,\zeta_i^*} \frac{1}{2}\|w\|^2 + c\sum_{i=1}^{N}\left(\zeta_i + \zeta_i^*\right) \qquad (7)$$

$$v_i - (w.\phi(u_i) + b) \le \varepsilon + \zeta_i$$

Subject to

$$\left(w.\phi(u_i) + b\right) - v_i \le \varepsilon + \zeta_i^*$$

$$\zeta_i, \zeta_i^* \ge 0, i = 1.2, \ldots .n$$

where w is the vector of coefficients ζ_i and ζ_i^* are the distances of the training data set points from the region where the errors less than ε are ignored and b is a constant. The index i label the ‘n’ training cases. Then, the dual form of the nonlinear SVR can be expressed as Eq. (8):

$$\max_{\underline{\alpha}_i . \alpha_i}$$

(8)

$$-\frac{1}{2}\sum_{i,j-1}^{n}(\underline{\alpha}_j - \overline{\alpha}_i)\left(\underline{\alpha}_j - \overline{\alpha}_i\right)\left\langle\phi(u_i).\phi u_j\right\rangle - \varepsilon\sum_{i=1}^{n}\left(\underline{\alpha}_j + \overline{\alpha}_i\right) + \sum_{i=1}^{n} v_i\left(\underline{\alpha}_j - \overline{\alpha}_i\right)$$

Subject to $\displaystyle\sum_{i=1}^{n}\left(\underline{\alpha}_i - \overline{\alpha}_i\right) = 0$ (8)

$$0 \le \underline{\alpha}_i \le c, i = 1, 2, \ldots, n \;\; 0 \le \overline{\alpha}_i \le c, i = 1, 2, \ldots, n$$

The “kernel trick” $k\left(u_i, u_j\right) = \left\langle \phi(u_i), \phi(u_j)\right\rangle$ is used for computations in input space χ to fetch the inner products into feature space $\Im$. Any function satisfying Mercer’s theorem should be used as kernels. Finally, the decision function of nonlinear SVR with the allowance of the kernel trick is expressed as Eq. (9):

$$g(u) = \sum_{i=1}^{n}(\alpha_i - \overline{\alpha}_i)k\left\langle u_i\, u\right\rangle + b \qquad (9)$$

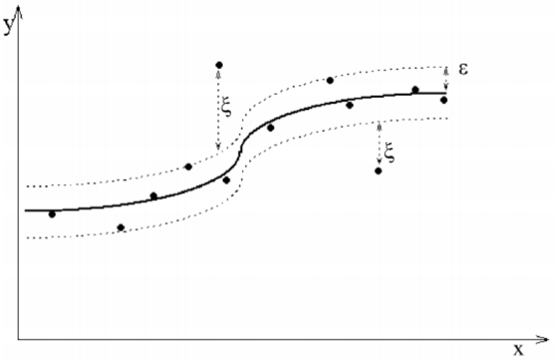

Figure 2) Nonlinear SVR with ε-insensitive loss function (It shows an example of a one-dimensional regression function with an ε-insensitive band. The variables ξ measure the cost of the errors on the training points) (Fletcher, 1987).

The parameters that impact over the effectiveness the nonlinear SVR are the cost constant C, the radius of the insensitive tube ε, and the kernel parameters. These parameters are mutually dependent over one another and hence altering the value of one parameter affects the other linked parameters also. The parameter C checks for the smoothness/flatness of the approximation function. A smaller value of C yields a learning machine with poor approximation due to under fitting of training data. A greater C value over fits the training data and sets its objective to minimize only the empirical risk making way for more complex learning. The parameter ε is related with smoothing the complexity of the approximation function and controls the width of the ε-insensitive zone used for fitting the training data. The parameter ε influences over the number of support vectors, and then both the complexity and the generalization capability of the approximation function is dependent upon its value. It also governs the precision of the approximation function. Smaller values of ε lead to more number of support vector and results in complex learning machine. Greater ε values result in more flat estimates of the regression function.

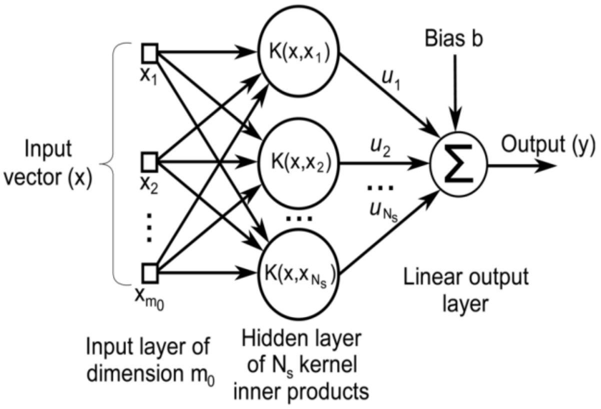

Figure 3) Network architecture of SVM (It shows the stages involved in the implementation of kernel pattern analysis. The data is processed using a kernel to create a kernel matrix, which in turn is processed by a pattern analysis algorithm to produce a pattern function) (Fletcher, 1987).

Determining appropriate values of C and ε is often a heuristic trial-and-error process. Figure3 shows the general network architecture of SVM.

3- Results

3.1- Hydrochemistry of groundwater

The mean concentrations of the major ions in the groundwater of Arak city are within the Iran Standard guidelines (TTPW, 2011) for drinking water (Table1). In the groundwater of Arak aquifers, concentrations of the Pb, Zn and Cu are higher than the recommended Iran Standard guidelines. The value of Pb, Zn and Cu ranges from 3 to 9 mg/l, 4 to 50 mg/l and 2 to 52 mg/l in the groundwater and the recommended Iran Standard ranges 0.05mg/l, 5 mg/l and 0.05 mg/l, respectively (Table 1).

Table1) Statistical characteristics of hydro-chemical variables in groundwater(units to mg/l).

Variables	Mean	Median	Minimum	Maximum
Cl	95	79	6.50	242
SO_4	213	245	23	320
HCO_3	166	156	55	410
Ca	242	250	80	400
Mg	22	20	7.50	43
Na	215	205	38	400
K	0.88	0.82	0.30	1.90
Fe	0.02	0.02	0.01	0.23
Mn	0.01	0.01	0.001	0.10
Pb	7.10	7	3	9
Zn	16	14	4	50
Cu	14	12	2	52

The maximum Cl and SO_4 concentrations of 242 mg/l and 320 mg/l respectively are, however higher than their respective Iran standard guidelines of 200 mg/l, and 250 mg/l. These are resulted from contamination of sources such as domestic sewage and agricultural activities. Maximum concentrations of some of the major ions such as Na are higher than the Iran standard. All other major parameters have concentrations lower than the standard guideline limits. The aquifers of the alluvial Arak, which are mostly sedimentary aquifers, therefore produce groundwater of acceptable quality for most uses.

3.2- Estimation of heavy metals using SVM-based regression

To simulate heavy metals in groundwater using SVR, all relevant parameters should be determined, due to the fact that (SVR) work based on given data and do not have previous knowledge about the subject of prediction. Following sections describe the input and output parameters and simulation of heavy metals in groundwater using SVR.

3.2.1- Input and output data

According to the correlation matrix HCO_3 and SO_4 that have most dependent on heavy metals (Pb, Zn and Cu) concentrations were selected as inputs of the network (Table 2).The outputs of network were heavy metals concentrations including Pb, Zn and Cu. In SVM-based regression any type of input can be used as long as they have effects on output results. To train and verify the accuracy and ability of the SVR, a total of 150 data samples records in groundwater from Arak city, were used in this research. In total, two input parameters including HCO_3,SO_4 (major ions) and output including Pb, Zn and Cu (heavy metals) were used to estimation of heavy metals in groundwater from Arak city.

3.2.2- Pre-processing of data

In data-driven system modeling methods, some pre-processing steps are usually implemented prior to any calculations, to eliminate any outliers, missing values or bad data. This step confirms that the raw data retrieved from database is perfectly proper for modeling. In order to softening the training procedure and improving the accuracy of prediction, all data samples are normalized to adapt to the interval [0, 1] according to the following linear mapping function (Eq. (10)):

$$x_M = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{10}$$

Table 2) Correlation matrix between heavy metals concentrations and independent variables.

	Cl	SO_4	HCO_3	Ca	Mg	Na	K	Fe	Mn	Pb	Zn	Cu
Cl	1.00											
SO_4	0.29	1.00										
HCO_3	-0.07	0.11	1.00									
Ca	0.35	0.56	0.36	1.00								
Mg	0.39	0.04	-0.01	0.19	1.00							
Na	0.49	0.68	0.42	0.39	0.00	1.00						
K	0.31	0.41	0.21	0.24	0.09	0.60	1.00					
Fe	0.05	-0.04	0.02	-0.11	0.03	0.13	0.05	1.00				
Mn	-0.05	0.05	0.15	0.06	0.08	0.04	0.09	0.07	1.00			
Pb	0.28	***0.85***	0.16	0.39	0.05	0.66	0.42	-0.07	0.07	1.00		
Zn	-0.03	0.16	***0.80***	0.36	0.11	0.35	0.19	0.01	0.19	0.19	1.00	
Cu	-0.08	0.15	***0.74***	0.35	0.23	0.28	0.14	-0.01	0.03	0.19	0.87	1.00

Where x is the original value from the dataset, x_M is the mapped value, and x_{min} (x_{max}) denotes the minimum (maximum) raw input values, respectively. It is to be noted that model outputs will be remapped to their corresponding real values by the inverse mapping function ahead of calculating any performance criterion. In this section, 80% of the datasets (120 samples) were assigned for training purposes arbitrary, while 20% (30 samples) was used for testing the network performance.

3.3- Modeling and performance criteria

Modeling was done in the statistica10 software. The main aim of this study was to build SVM models for the regression problems pertaining to the groundwater quality with a view to develop a tool for the prediction of heavy metals using simple and directly measurable water quality parameters as the input. Similar to other multivariate calibration methods, the generalization performance of SVM regression models depends on a proper setting of several parameters. These include the capacity parameter C, the insensitive loss function ε, and the kernel function dependent parameter in SVM regression models (Khan and Coulibaly, 2006). RBF is the most commonly used kernel in SVM and the RBF width parameter (γ) reflects the distribution/range of x-values of training data (Hazi *et al*., 2010). The parameter C determines the trade-off between the smoothness of the regression function and the amount up to which deviations larger than ε are tolerated. Therefore, the choice of the C value influences the significance of the individual data points in the training set (Hazi *et al*., 2010). Hence, a proper choice of C in combination with ε might result in a well performing and robust regression model, which is also insensitive to the presence of possible outliers. Here, the optimum value of C was determined through grid search over a space of 0.01–50,000.A good combination of the two parameters (C and ε) also prevents overtraining. To achieve this, an internal cross-validation during construction of SVR models was performed. The kernel function is used to map the input data into a high dimensional feature space which is required to transform the nonlinear input space to a high-dimensional feature space where linear regression is possible. The mapping depends on the intrinsic structure of the data, implying that the kernel type and parameters need be optimized to approximate the ideal mapping (Bray and Han, 2004). In this work, RBF kernel was used. Unlike the linear kernel, the RBF kernel can handle the case when the relation between attributes is nonlinear. It is also worth mentioning that the RBF kernel is good in cases where the input dimension is low, as it projects the data to a higher latent dimensionality. The linear kernel is good in situations where the input dimension is already high and has many null values. Besides, the linear kernel is a special case of the RBF. The RBF kernel has fewer tuning parameters than the polynomial and sigmoid kernels and it tends to give good performance under general smoothness assumptions (Chen and Yu, 2007). Here, the optimum value of the RBF kernel function (γ) was determined through the grid search over the space 0.001–20.

To evaluate the performances of the SVR and MLR model, root mean squared error (RMSE), correlation coefficient (R) and variance account for (VAF) were chosen to be the measure of accuracy. Let yk be the actual value and be the predicted value of the kth observation and n be the number of samples. The higher the R and VAF the better is the model performance. For instance, VAF of 100% means that the measured output has been predicted exactly (perfect model). It can also mean that the model is over fitting. R and VAF =0 means that the model performs as poorly as a predictor using simply the mean value of the data. Also, the lower RMSE indicates the better performance of the model. RMSE, R and could be defined, respectively, as Eqs. (11)-(12)-(13):

$$RMSE = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(y_k - \hat{y}_k)^2} \tag{11}$$

$$R = \sqrt{\frac{(\sum_{k=1}^{n} y_k \hat{y}_k - n\mu_y \mu_{\hat{y}})^2}{(\sum_{k=1}^{n} \hat{y}_k^2 - n\mu_y{}^2)^2 (\sum_{k=1}^{n} \hat{y}_k^2 - n\mu_{\hat{y}}^2)^2}} \tag{12}$$

$$\mathrm{VAF} = \left(1 - \frac{\mathrm{Var(yk}- \hat{y}_k}{\mathrm{Var(yk)}}\right).100\% \tag{13}$$

Where $\mu_y\ (\mu_{\hat{y}})$ denotes the mean value of the $\mu_k\ (\mu_{\hat{k}}),\ k=1,...,n,$ respectively and var. denotes the variance.

4- Discussion

The SVR approach was used for predicting the heavy metals of groundwater using a set of simple and directly measurable water quality variables. The complete water quality data set was divided in to two sub-sets (training and test). In SVR, heavy metals was the dependent variable, whereas, the HCO_3, SO_4 variables constituted the set of independent variables. Among the linear, polynomial, sigmoid, and RBF kernel functions, the later was finally selected in SVR models as it yielded the highest R. Moreover, the RBF kernels tend to give good performance under general smoothness assumption (Chen and Yu, 2007; Wei *et al*., 2007). The values of the model performance criteria parameter (R) as computed for the training and test sets used for the model are presented in Table 3. For the heavy metal values predicted by the model, the correlation coefficient (R) values ($p < 0.001$) for the training and test sets were 0.86,0.81 for Pb,0.77,0.91 for Zn and 0.68,0.87 for Cu in the RBF model respectively. The SVR predictions are precise, if R values are closer to unity.

Table 3) Values of the performance criteria parameter for SVR models

		Models	R	MSE	VAF%
Pb	Training	RBF	0.86	0.123	85
		Polynomial	0.78	0.168	74
		Sigmoid	0.38	0.199	42
		Linear	0.86	0.123	85
	Testing	RBF	0.81	0.158	81
		Polynomial	0.77	0.179	73
		Sigmoid	0.64	0.186	62
		Linear	0.80	0.144	79
Zn	Training	RBF	0.77	0.124	73
		Polynomial	0.76	0.176	72
		Sigmoid	-0.02	0.341	0.03
		Linear	0.74	0.131	68
	Testing	RBF	0.91	0.121	89
		Polynomial	0.88	0.119	87
		Sigmoid	-0.22	0.226	0.24
		Linear	0.91	0.108	89
Cu	Training	RBF	0.68	0.198	65
		Polynomial	0.63	0.182	60
		Sigmoid	-0.04	0.321	0.06
		Linear	0.67	0.191	64
	Testing	RBF	0.87	0.128	86
		Polynomial	0.83	0.141	75
		Sigmoid	-0.35	0.296	26
		Linear	0.78	0.168	74

The performance indices obtained in Table 3 indicate the high performance of the SVR model that can be used successfully for the estimation of heavy metals in the groundwater. Furthermore, correlation between measured and predicted values of heavy metals in the groundwater for training and testing phases are shown in Figures 4 and 5. In order to increase the accuracy and applicability of SVR for estimation of heavy metals in groundwater, SVR algorithm was used to weighting SVR. Several SVR models were trained and tested using obtained data from Arak city, to determine the optimum network. Performances of the selected SVR model using training and testing dataset are shown in Figures 4 and 5 and Table 3.

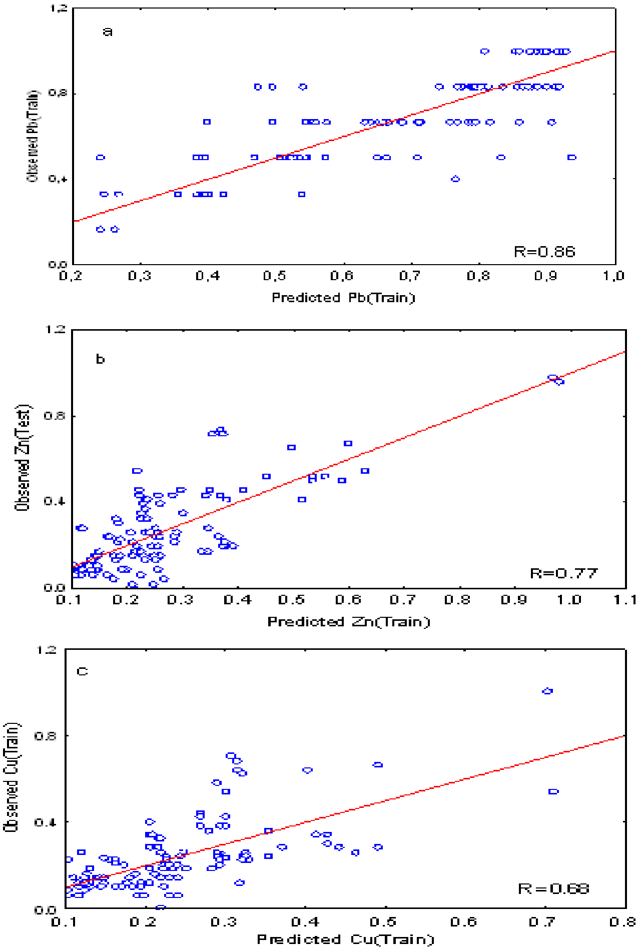

Figure 4) Correlation between measured and predicted values of heavy metals in the groundwater for training data sets, a) Pb, b) Zn, c) Cu.

The predicted heavy metals fit the measured heavy metals almost perfectly for training datasets. Nevertheless, the predicted heavy metals denote fit perfectly to the measured heavy metals for testing datasets. This might be caused by a lack of training data in that range. In general, it can be said that the proposed SVR model is able to predict heavy metals with high degree of accuracy. Table 4 compares the correlation coefficient R associated with both training and test data.

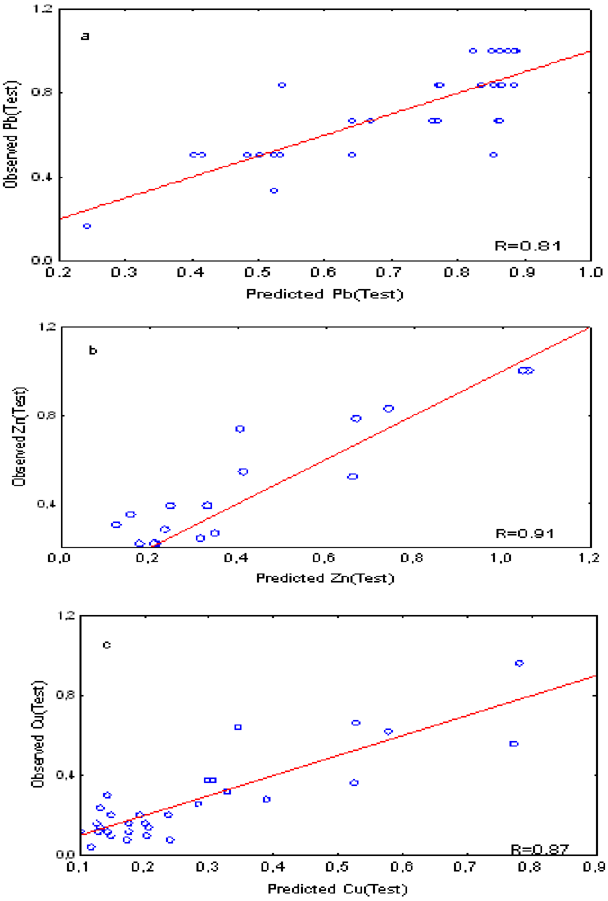

Figure 5) Correlation between measured and predicted values of heavy metals in the groundwater for testing data sets, a) Pb, b) Zn, c) Cu.

Table 4) The comparison of the results (R) of training and test data.

	Pb		Zn		Cu	
	Train	Test	Train	Test	Train	Test
RBF	0.86	0.81	0.77	0.91	0.68	0.87
Polynomial	0.78	0.77	0.76	0.88	0.63	0.83
Sigmoid	0.38	0.64	-0.02	- 0.22	-0.04	- 0.35
Linear	0.86	0.80	0.74	0.91	0.67	0.87

5- Conclusion

High concentrations of Pb, Zn and Cu were found in the groundwater of Arak City. Heavy metals were emitted mainly by anthropogenic sources. In this paper, SVM Regression model was developed to estimation of heavy metals in the groundwater from Arak city, Iran. To generate the proposed SVR model, a dataset consists of 150 samples was used. Two parameters including HCO3, SO4, and (major ions) were used as input parameters and Pb, Zn and Cu (heavy metals) were used as output parameters. Consequently, it may conclude that SVR is a reliable system modeling technique for estimation of heavy metals in the groundwater from Arak city with highly acceptable degree of accuracy and robustness.

Acknowledgements

Chemical analyses were conducted at the laboratory of the Department of Mining Engineering, Arak University of Technology. The authors thank Arak University of Technology for their continuous support.

References

Abbasi, M., Abdulli, M. A., Omidvar, O., Baghvand, A. 2013. Forecasting Municipal Solid waste Generation by Hybrid Support Vector Machine and Partial Least Square Model. International Journal of Environmental Research: 7, 27–38.

Aguirre-Basurko, E., Ibarra-Berastegi, G., Madariaga, I. 2006. Regression and multilayer perceptron-based models to forecast hourly O_3 and NO_2 levels in the Bilbao area. Environment Modeling Software 21: 430–446.

Anderson, R. B., Naftz, D. L., Day-Lewis, F. D., Henderson, R. D., Rosenberry, D. O., Stolp, B. J., Jewell, P. 2014. Quantity and quality of groundwater discharge in a hypersaline lake environment. Journal of Hydrology 512: 177–194.

Andre, L., Franceschi, M., Pouchan, P., Atteia, O. 2005. Using geochemical data and modeling to enhance the understanding of groundwater flow in a regional deep aquifer, Aquitaine Basin, south-west of France. Journal of Hydrology: 305, 40–62.

Bishop, C.M. 2006. Pattern Recognition and Machine Learning, Springer, New York, 330p.

Bray, M., Han, D. 2004. Identification of support vector machines for runoff modeling. Journal of Hydro informatics: 6, 265–280.

Chen, S. T., Yu, P. S. 2007. Pruning of support vector networks on flood forecasting. Journal of Hydrology: 347, 67–78.

Chen, S. T. 2015. Mining Informative Hydrologic Data by Using Support Vector Machines and Elucidating Mined Data according to Information Entropy. Entropy: 17, 1023 –1041.

Deng, N., Tian, Y., Zhang, C. 2012. Support Vector Machines: Optimization Based Theory, Algorithms, and Extensions. Chapman and Hall/CRC, New York.112p.

Devic, G., Djordjevic, D., Sakan, S. 2014. Natural and anthropogenic factors affecting the groundwater quality in Serbia. Science of the Total Environment: 468–469, 933–942.

Fletcher, R. 1987. In: Practical Methods of Optimization, second ed., Wiley, New York, 400p.

Hassan, M., Shamim, M. A., Hashmi, H. N., Ashiq, S. Z., Ahmed, I., Pasha, G. A., Naeem, U. A., Ghumman, A. R., Han, D. 2014. Predicting stream flows to a multipurpose reservoir using artificial neural networks and regression techniques. Earth Science Informatics: 8, 337–352.

Haykin, S. 1999. Neural Networks. Comprehensive Foundation, Prentice Hall, New Jersey, 550p.

Helstrup, T., Jorgensen, N., Banoeng-Yakubo, B. 2007. Investigation of hydrochemical characteristics of groundwater from Cretaceous–Eocene limestone aquifers in southern Ghana and Togo using hierarchical cluster analysis. Hydrogeology: 15, 977–989.

Hazi, M. A., Aminuddin, A, G., Chang, K. C., Zorkeflee, A. H., Zakaria, A. 2010. Machine learning approach to predict sediment load– a case study. Journal of Clean Soil, Air, Water: 38, 969–976.

Khan, S. M., Coulibaly, P. 2006. Application of support vector machine in lake water level prediction. Journal of Hydrologic Engineering ASCE: 11, 199–205.

Masoud, A. A. 2014. Groundwater quality assessment of the shallow aquifers west of the Nile Delta (Egypt) using multivariate statistical and geostatistical techniques. Journal of African Earth Sciences: 9, 123–137.

Matatos, I., Alexopoulos, A., Godelitsas, A. 2014. Multivariate statistical analysis of the hydrogeochemical and isotopic composition of the groundwater resources in northeastern Peloponnesus (Greece). Science of the Total Environment: 476–477, 577–590.

Monjerezi, M., Vogt, M. R., Aagaard, P., Saka, J. D. K. 2011. Hdro-geochemical processes in an area with saline groundwater in lower Shire River valley, Malawi: An integrated application of hierarchical cluster and principal component analyses. Applied Geochemistry: 26, 1399–1413.

Oinam, J. D., Ramanathan, A. L., Singh, G. 2012. Geochemical and statistical evaluation of groundwater in Imphal and Thoubal district of Manipur. India Journal of Asian Earth Sciences: 48, 136–149.

Shawe-Taylor, J., Cristianini, N. 2004. Kernel Methods for Pattern Analysis, Cambridge University Press, New York, 265p.

Singh, K. P., Gupta, S., Mohan, D. 2014. Evaluating influences of seasonal variations and anthropogenic activities on alluvial groundwater hydrochemistry using ensemble learning approaches. Journal of Hydrology: 511, 254–266.

Srinivasamoorthy, K., Chidambaram, S., Vasanthavigar, M., Sarma, V. S. 2014. Hydrochemical characterization and quality appraisal of groundwater from Pungar sub basin, Tamilnadu, India. Journal of King Saud University-Science: 26, 37–52.

Suárez Sánchez, A., García Nieto, P. J., Riesgo, Fernández, P., Coz Díaz, J. J., Iglesias-Rodríguez, F. J. 2011a. Application of a SVM-based regression model to the air quality study at local scale in the Avilés urban area (Spain). Mathematic Computer Modeling: 54, 1453–1466.

Suárez Sánchez, A., Riesgo Fernández, P., Sánchez Lasheras, F., Cos Juez, F. J., García Nieto, P. J. 2011b. Prediction of work-related accidents according to working conditions using support vector machines. Applied Mathematic Computer: 218, 539–3552.

TTPW. 2011. Tehran Province Water and Wastewater. Water standard of Iran.

Wang, P., Yu, J., Zhang, Y., Liu, C. 2013. Groundwater recharge and hydrogeochemical evolution in the Ejina Basin, northwest China. Journal of Hydrology: 476, 72–86.

Wei, W., Wang, X., Xie, D., Liu, H. 2007. Soil water content forecasting by supportvector machine in purple hilly region. Computer and computing technologies in agriculture, International Federation for Information Processing: 258, 223–230.

Yidana, S. M. 2010. Groundwater classification using multivariate statistical methods: Southern Ghana. Journal of African Earth Sciences: 57, 455–469.

Zamani, F. 1999. Sedimentology of Arak Mighan lake. Msc Thesis in Beheshti University, 225p.

Zapata, E. P., Ruiz, R. L., Harter, T., Ramírez, A. I., Mahlknecht, J. 2014. Assessment of sources and fate of nitrate in shallow groundwater of an agricultural area by using a multi-tracer approach. Science of The Total Environment: 470–471, 855–864.

Zghibi, A., Merzougui, A., Zouhri, L., Tarhouni, J. 2014. Understanding groundwater chemistry using multivariate statistics techniques to the study of contamination in the Korba unconfined aquifer system of Cap-Bon (North-East of Tunisia). Journal of African Earth Sciences: 89, 1–15.