Environmental Resources Research
Vol. 6, No. 2, 2018

GUASNR

Comparison and evaluation of intelligent models for river suspended sediment estimation (case study: Kakareza River, Iran)

H. Torabi[1], R. Dehghani[2*]
[1]Associte Professor of Water Engineering, Lorestan University, Khorramabad, Iran
[2]Ph.D. student of Water Structure, Faculty of Agricultural, Lorestan University, Khorramabad, Iran

Received: August 2017 ; Accepted: September2017

Abstract

Sediment transport constantly influences river and civil structures and the lack of information about its exact amount makes management efforts less effective. Hence, achieving a proper procedure to estimate the sediment load in rivers is important. We used support vector machine model to estimate the sediments of the Kakareza River in Lorestan Province and the results were compared with those obtained by gene expression programming. The parameter of flow discharge for input in different time lags and the parameter of sediment for output during 1992-2012 were considered. Criteria including correlation coefficient, root mean square error and mean absolute error were used to evaluate and also compare the performance of models. With regards to accuracy, the support vector machine model showed the highest correlation coefficient (0.994), minimum root mean square error (0.001 ton/day) and the mean absolute error (0.001 ton/day) which was initiated at verification stage. The results also showed that the support vector machine has great capability to estimate the minimum and maximum values for sediment discharge.

***Keywords*:** Suspended sediment, Kakareza, Support vector machine, Gene expression programing

*Corresponding author; reza.dehghani67@yahoo.com

Introduction

Historically, there have been a number of attempts to estimate sediment yield in rivers using modeling that can be broken down into different groups (White, 2005). The deterministic models can be grouped by either empirical or conceptual. These models generally need long data records and take into account the hydrodynamics of each mode of transport. The deterministic and stochastic models are based on the physical processes of the sediment yield, some of which have been reported in the literature (Singh *et al*., 1998; Yang, 1996; Cohn *et al*., 1992; Forman *et al*., 2000) for sediment discharge estimation. The application of the physics-based software programs necessitates detailed spatial and temporal environmental data that is not often available. In practice, the most commonly used model is the rating curve model, which is based on the relationship between flow (Q) and sediment (S). The amount of sediment yield in a river is measured as sediment load (S), which depends upon the sediment concentration and the river discharge (Q). Accurate estimation of the sediment yield is rather difficult because of the temporal variation of both the sediment concentration and the river discharge.

Generally, the time-series techniques consider linear relationships among variables. However, these techniques are difficult to employ for the real hydrologic data due to the temporal variations. In contrast, support vector machine (SVM) model is a nonlinear model and can be used to identify these relations. Neural networks are increasingly being used in diverse engineering applications because of their ability to solve nonlinear regression problems successfully. This feature is one of the highly important aspects of neural computing, because it allows to model a function with little information or incomplete understanding. Thus, the SVM approach is extensively used in water resources literature for prediction and forecasting. In recent years, SVM has been widely used in various fields. Runoff and sediment yield estimation can utilize SVM as well (Misra *et al.,* 2009) which is a powerful nonlinear pattern recognition technique (Vapnik, 1998; Kecman, 2000). In this study, we estimated suspended sediment load using linear regression model, power regression model, artificial neural network and support vector machine. Records of river discharge and suspended sediment loads in Kaoping river basin were investigated as a case study and the result showed that SVM outperforms the ANN and other two regression models (Chiang and Tsai, 2011). Gene-expression programming (GEP), which is an extension of genetic programming (GP), as an alternative approach for modeling the functional relationships of sediment transport in sewer pipe systems was studied satisfactorily by Ghani and Azamathulla, (2011). The study of river discharges and suspended sediment loads in the Goodwin Creek Experimental Watershed in United States used a similar methodology. As a result, we believe that the proposed SVM model has high potential for predicting suspended sediment load (Chiang *et al.,* 2014). Another study compared the results of the Soil and Water Assessment Tool (SWAT) with a Support Vector Machine (SVM) to predict the monthly streamflow of an arid region located in the southern part of Iran, namely the Roodan Watershed. The results indicated that the SVM had a closer value for the average flow in comparison to the SWAT model; whereas the SWAT model outperformed for total runoff volume with a lower error in the validation period (Jajarmizadeh *et al.,* 2015). Discharge time series were investigated using predictive models of support vector machine (SVM) and artificial neural network (ANN) and their performances were compared with two conventional models. The evaluation of the results showed different performance measures, which indicated that SVM and ANN had an edge over the results by the conventional models. Notably, peak values predicted by SVM and ANN were more reliable than the conventional models, although the performances of these latter were acceptable for a range of practical problems (Ghorbani *et al.*, 2016). The purpose of this research is estimation of

suspended sediment in Kakareza River using support vector machine and comparison of its results with gene expression programming.

Materials and Methods

Case study and used data

The study area is Kakareza River in Lorestan Province, Iran. This river is one of the permanent rivers in the province, and is originated from southeastern mountains of Aleshtar and Biranshahr (Dehno). When this river passes through Aleshtar suburbs, it is known as Kakareza. The river is between "15° 48° 49° longitude and " 22° 32 to "52° 33 degrees latitude and it flows across the east of Khorramabad (capital city of Lorestan Province). This river is one of the initial branches of Karkhe River in Zagros mountains and have an average altitude of 1550 meters above sea level. Karkhe River basin area is about 1148 square kilometers and its river has a length of 85 km. Kakareza joins Kashkan, Cimmeria, and Karkhe Rivers in its way and eventually pours into the Persian Gulf. The geographical location of the study area is shown in Figure 1. In this study, available sediment data at monthly scale of Horod station (Kakareza) from 1992 to 2012 was used provided by Lorestan Regional Water authority. Table 1 shows the statistical properties of Kakareza River during the mentioned period.

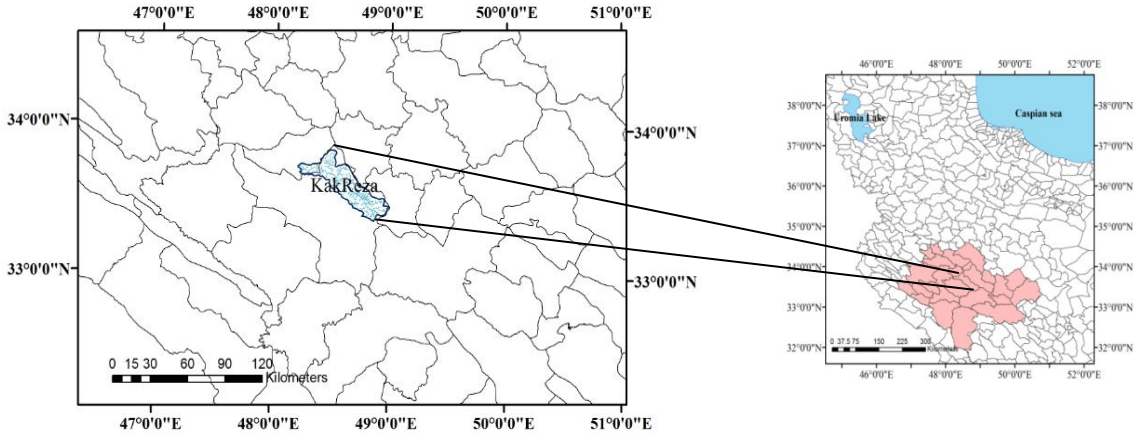

Figure 1. Geographical location of Kakareza River

Table 1. Statistical properties sediment parameter during 1992-2012

Station	Period of record	Data set	Statistics	Q(m3/s)	S(ton/day)
Kakareza	1992-2012	Training	Minimum	0.780	0.324
			Mean	16.575	3822.239
			Maximum	370.740	285988.254
			deviation	38.551	24361.715
			Skewness	5.703	9.386
		Testing	Minimum	0.000	0.669
			Mean	6.078	27.692
			Maximum	20.230	191.101
			deviation	5.894	48.117
			Skewness	1.046	2.540

Gene Expression Programming

Gene Expression Programming method was presented by Ferreira in 1999 (Ferreira, 2001). This method is a combination of genetic algorithm (GA) and genetic programming (GP) method in which, simple linear chromosomes of fixed length are similar to what is used in genetic algorithm and branched structures with different sizes and shapes are similar to the decomposition of trees in genetic programming. Since in this method all branch structures of different shapes and size are encoded in linear chromosome with

fixed length, the system could use all evolutionary advantages of the two approaches. In this method different phenomena are modeled by collection of functions and terminals. Collection of functions consists of the main arithmetic functions {+, -, ×, /}, the trigonometric functions or any other mathematical function {√, x2, sin, cos, log, exp, …} or can be defined functions by user as deemed appropriate for modeling. Collection of terminals consists of constant values of problem and independent variables (Ferreira, 2001). For applying gene expression programming method Gen Xpro Tools 4.0 Software was used (Ghorbani *et al*., 2012).

Support Vector Machine

Support Vector Machine is an efficient learning system based on optimization theory that uses the principle of minimization of structural error and results in an overall optimal solution (Vapnik, 1998). In regression model, SVM applies a function with the dependent variable Y for several independent variables X (Xu *et al*., 2007). Like other regression problems the method assumes the relationship between the dependent and independent variables to be determined with algebraic functions similar to f(x) plus some allowable error (ε).

$$f(x)=W^{T}.\emptyset(x)+b \quad (1)$$

$$y=f(x)+noise \quad (2)$$

where W is coefficients vector, b is constant of regression function, and $\emptyset$ is kernel function, and the goal is to find a functional form for f(x). This is realized with SVM model training using collection of samples (train collection). To calculate w and b requires error optimization function in ε-SVM considering the conditions embodied in Equation 4 (Shin *et al*., 2005).

$$W^{T}.\emptyset(X_i)+b-y_i \le \varepsilon+\varepsilon_i^{*} \;,\; \frac{1}{2}W^{T}.W + C\sum_{i=1}^{N}\varepsilon_i + C\sum_{i=1}^{N}\varepsilon_i^{*} \quad (3)$$

$$y_i - W^{T}.\emptyset(X_i) - b \le \varepsilon + \varepsilon_i \;,\; \varepsilon_i, \varepsilon_i^{*} \ge 0 \;,\; i=1,2,\dots,N \quad (4)$$

In the above equations, C is integer and positive, and it's factor of penalty determinant when an error occurs. $\emptyset$ is kernel function, N is number of samples and ε_i and ε_i^{*} are shortage variables. Finally, we can rewrite SVM function as follows (Shin *et al*., 2005):

$$f(x)=\sum_{i=1}^{N}\bar{\alpha}_i\,\emptyset(x_i)^{T}.\emptyset(x)+b \quad (5)$$

Average Lagrange Coefficients $\bar{\alpha}_i$ in characterized space is $\emptyset(x)$. To simplify the formula, the usual process of SVM model selects a kernel function as follows.

$$K(X_J,X)=\emptyset(X_i)^{T}\sqrt{b^2-4ac} \quad (6)$$

Different kernel functions can be used to create different types of ε-SVM. Various kernel functions used in SVM regression models are: Polynomial with three characteristics of the target, Radial Basis Functions (RBF) with one characteristics of the target, and Linear method which respectively are shown below (Vapnik, 1998).

$$k(x_i,x_j)=(x_i.x_j)^{d} \quad (7)$$

$$K(x,x_i)=\exp\left(-\frac{\|x-x_i\|^2}{2\sigma^2}\right) \quad (8)$$

$$k(x_i,x_j)=x_i.x_j \quad (9)$$

Evaluation Criteria

In this research, to evaluate the accuracy and efficiency of the models we used indices Correlation Coefficient (CC), Root Mean Square Error (RMSE), Nash–Sutcliffe coefficient (NS), and Bias according to the following relations. Best values for these four criteria are respectively 1, 0, 1, and 0.

$$CC=\frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i-\bar{x})^2\sum_{i=1}^{N}(y_i-\bar{y})^2}} \quad -1\le R\le 1 \quad (10)$$

$$RMSE=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-y_i)^2} \quad (11)$$

$$NS=1-\frac{\sum_{i=1}^{N}(x_i-y_i)^2}{\sum_{i=1}^{N}(x_i-\bar{y})^2} \quad -\infty\le NS\le 1 \quad (12)$$

In the above relations x_i and y_i are respectively observed and calculated values in time step i, N is number of time steps, $\bar{x}$ and $\bar{y}$ are respectively mean observed and calculated values.

Results and Discussion

One of the most important steps in modeling is selection of the right combination of input variables shown in Table 2.

Table 2. The structure of input combinations

Structure	Input	Output
1	Q(t)	S(t)
2	Q(t)Q(t-1)	S(t)
3	Q(t)Q(t-1)Q(t-2)	S(t)

In this Table Q(t), Q(t-1) and Q(t-2) are respectively discharge in t, t-1 , t-2 time as input and S(t) is sediment in time t as output being considered. Due to the significant cross-correlation between input and output data, in order to achieve an optimal model to estimate the sediments in Kakareza River, different combinations of input parameters shown in Table 3 were used. To estimate input discharge in Kakareza River through GEP and SVM, we gathered 240 registered records of the catchment hydrometric data during the period 1992-2012, of which we used 192 for training and 48 for verification.

We selected important variables in the model and removed those with less impact in GEP which also provided the ability to have a clear relationship for estimating sediment in the Kakareza River. Results of gene expression programming model for both operators in Table3 showed that F2 operator has higher accuracy in both training and verification stages with correlation coefficient maximum R=0.813, root mean square error RMSE=0.002, mean absolute error MAE=0.002 and NS=0.643.

Also, in order to compare the results, we applied the support vector machine model in MATLAB (The MathWorks Inc., 2012). In this study the RBF, Poly and Line kernel with parameters (C, ε, σ), were used for stage–discharge modeling, with the accuracy of the SVM model being dependent on the identified parameters. The parameter search scheme was the shuffled complex evolution algorithm (SCE-UA), (Lin *et al*., 2006; Yu *et al*., 2006). The SCE-UA technique has been used successfully in the area of surface and subsurface hydrology processes (Duan *et al*., 1994). To obtain suitable values of these parameters (C, ε, σ), the RMSE was used to optimize parameters. In order to estimate the sediment in Kakareza River by SVM model, we examined kernel types of linear kernel, polynomial and radial basis functions that are commonly used in hydrology. The results are given in Table3. According to this table, the combined model number 3 as radial basis functions kernel has the highest correlation coefficient R=0.994, lowest root mean square error RMSE=0.001 ton/day, mean absolute error MAE=0.001ton/day and NS=0.988 in verification stage than other models. In Figure 3 the best model for verification of data is shown.

Table 3. The final results of the training and verification in gene expression programming and support vector machine

Model	Training				Testing			
	R	RMSE	MAE	NS	R	RMSE	MAE	NS
SVM_RBF_1	0.91	0.074	0.27	0.901	0.946	0.008	0.006	0.952
SVM_RBF_2	0.95	0.042	0.011	0.926	0.978	0.005	0.003	0.978
SVM_RBF_3	**0.974**	**0.018**	**0.006**	**0.945**	**0.994**	**0.001**	**0.001**	**0.988**
GEP_F2_1	0.89	0.075	0.023	0.837	0.797	0.011	0.007	0.612
GEP_F2_2	0.92	0.043	0.014	0.862	0.805	0.007	0.003	0.637
GEP_F2_3	0.936	0.030	0.008	0.876	0.813	0.002	0.002	0.643

Figure 2 shows the ability of SVM model in estimation of most values. The scatter plots of gene expression programming related to the verification stage in Fig (2-b) showed the fit line of computational values with four mathematical operators to the best fit line y=x. Based on this figure, all of the estimated and observation values are in the fit line except few points. In fig 3-a, SVM model has acceptable performance for estimation. But according to Fig 3-b, the GEP model has not been good in estimating the maximum value.

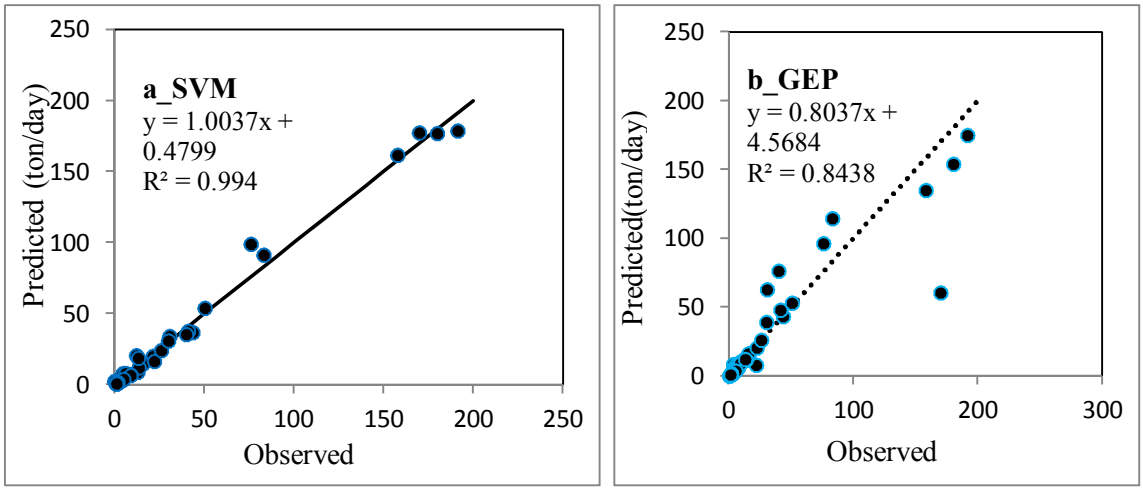

Figure 2. Scatterplots of the predicted-observed sediment time series of the Kakareza station in test period using (a) SVM; (b) GEP.

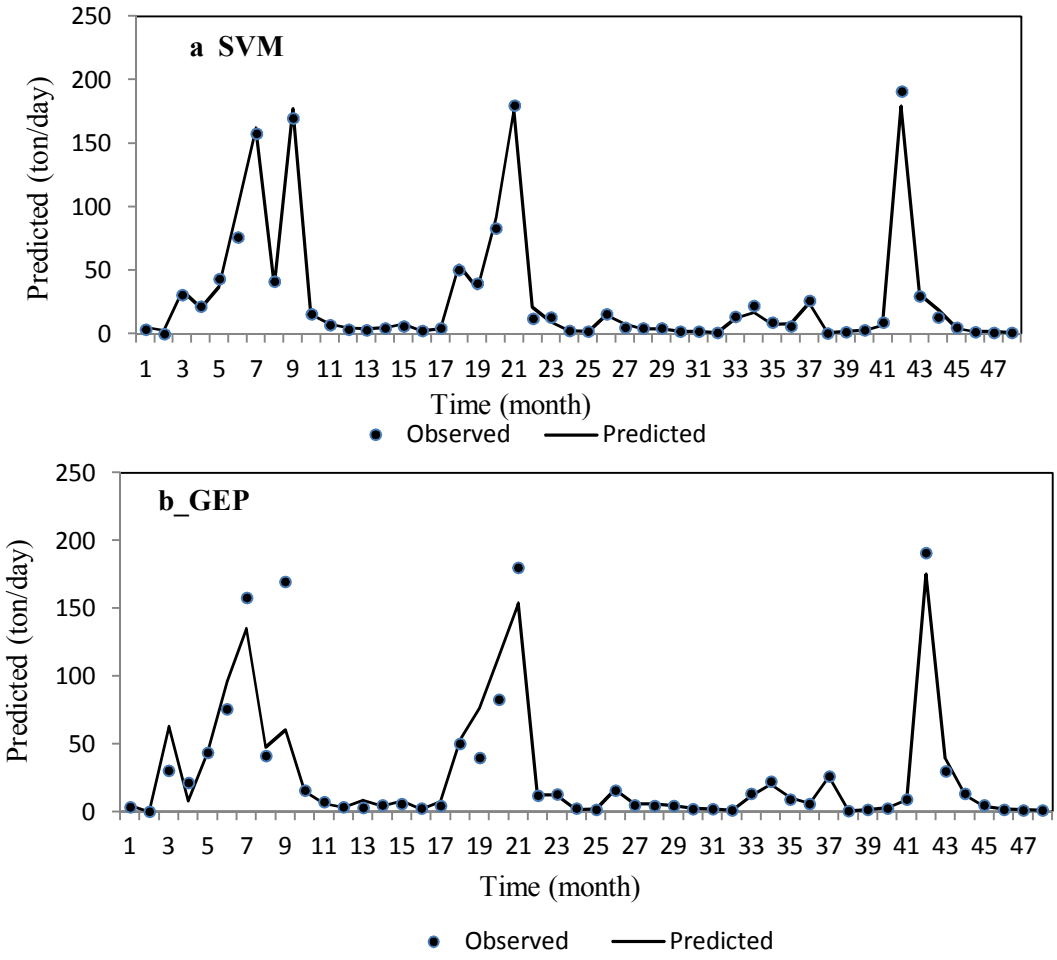

Figure 3. Comparison of the optimal models with observed values for testing the data set (a) SVM; (b) GEP

Performance comparison of the models

The two models provided good simulates of the sediment load in Kakareza River. Comparison of gene expression programming and support vector machine showed proximity of the two. In Figure 4 the estimated and observed values in gene expression programming and support vector

machine models for recorded data in verification stage is shown, in which the support vector machine model well approximates the minimum, maximum, and middle values.

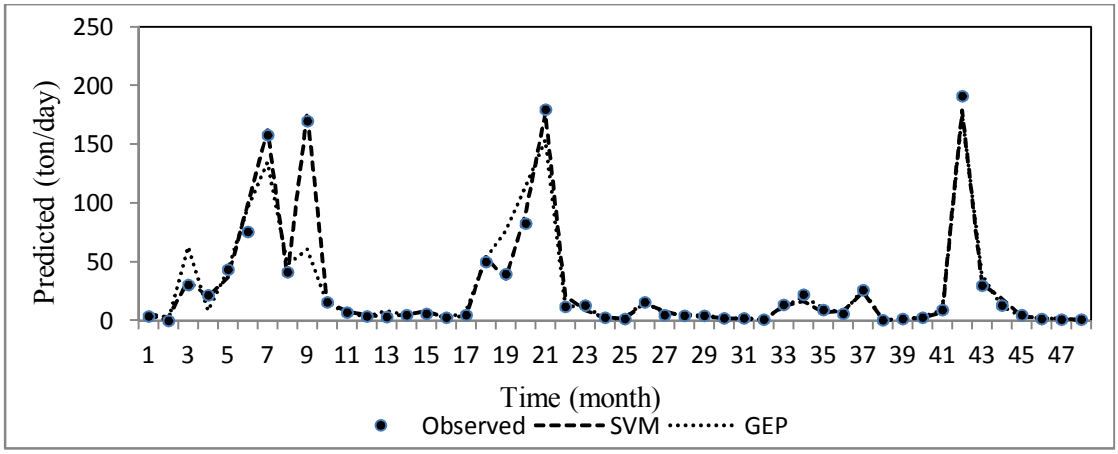

Figure 4. The scatter plot between estimated and observed values in gene expression programming and support vector machine models for recorded data in verification stage

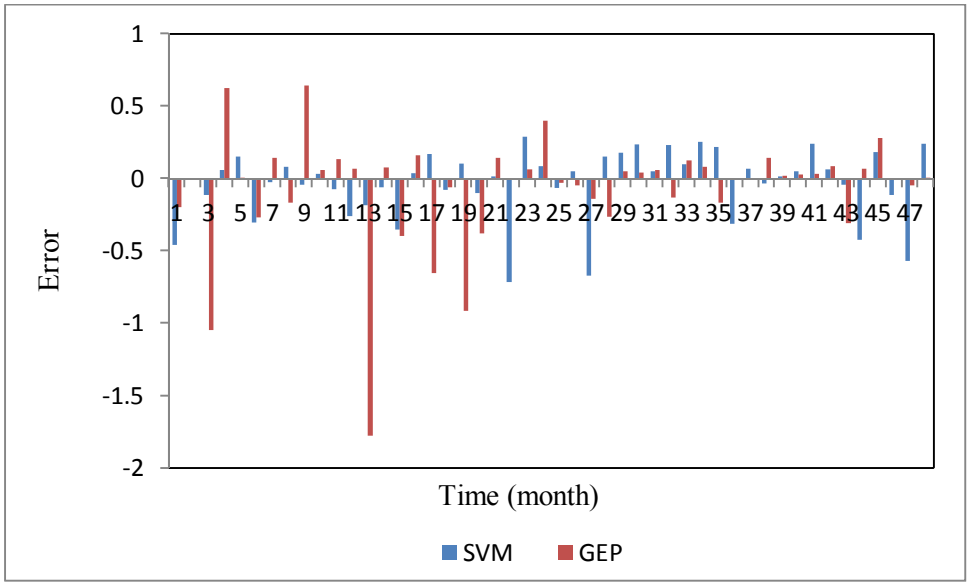

Figure 5. The two models optimization error graph as a percentage of the mean observed value

The difference between the observed sediment values and optimal computational models calculated as a percentage of the mean observed values (error value) in comparison with the recorded data in shown in Figure 5. Among these models, the SVM has the lowest error value. Due to the high estimation accuracy and reliability of gene expression programming and support vector machine models, correlation between the observed and the computed values are respectively 0.994 and 0.813. Also, SVM model has significant correlation in the probability levels %5 and %10.

Conclusions

In this research, we tried to evaluate the performance of gene expression programming and support vector machine for simulating sediment in the Kakareza River in Lorestan Province of Iran using sediment monthly data. The observed sediment values were compared with the

estimated sediment in these models (GEP and SVM). The results suggest that the SVM model has high accuracy and low error in estimating minimum, maximum, middle values and peak sediment. Also, the gene expression programming model with four basic arithmetic operations showed high ability to estimate minimum, maximum, and middle values and peak values like the support vector machine with radial basis functions kernel to estimate the minimum and middle values. Estimating sediment using combined models had lower error and high correlation than other models.

Totally, the results of this research showed that support vector machine method has the highest accuracy. This result is supported by the study of Jajarmizadeh *et al*. (2015) and Ghorbani *et al*. (2016). Also this research showed gene expression programming and support vector machine models could be used safely to estimate the sediments in Kakareza River.

Acknowledgments

The authors are very grateful to the Regional Water Company, Lorestan Province, Iran, for data gathering during this research.

References

Chiang, J., Tsai, Y., Cheng, K., Lee, Y., Sun, M., and Wei, J. 2014. Suspended Sediment Load Prediction Using Support Vector Machines in the Goodwin Creek Experimental Watershed. Geophysical Research Abstracts. 16(1), 234-247.

Chiang, J., Tsai, Y., 2011. Suspended Sediment Load Estimate Using Support Vector Machines in Kaoping River Basin. International Conference onsuspended sediment load

Cohn, T.A., Caulder, D.L., Gilroy, E.J., Zynjuk, L.D., and Summers, R.M. 1992. The validity of a simple statistical model for estimating fluvial constituent loads: an empirical study involving nutrient loads entering Chesapeake Bay. Water Resources Research 28, 2353–2363.

Ferreira, C. 2001. Gene expression programming: a new adaptive algorithm for solving problems. Complex Systems, 13(2), 87–129.

Forman, S.L., Pierson, J., and Lepper, K. 2000. Luminescence geochronology. In: Sowers, J.M., Noller, J.S., Lettis, W.R. (Eds.), Quaternary Geochronology: Methods and Applications. American Geophysical Union Reference Shelf 4, Washington DC, 157–176.

Ghani, A.B., and Azamathulla, H. 2011.Gene-Expression Programming for Sediment Transport in Sewer Pipe Systems. J. Pipeline Syst. Eng. Pract., 2(3), 102-106.

Ghorbani, M.A., Khatibi, R., Goel, A., and Azani, A. 2016. Modeling river discharge time series using support vector machine and artificial neural networks. Environmental Earth Sciences. 75(8), 675-685.

Ghorbani, M.A., Khatibi, R., Asadi, H., and Yousefi, P. 2012. Inter- Comparison of an Evolutionary Programming Model of Suspended Sediment Time-series whit other Local Model. INTECH. Pp, 255-282.

Ghorbani, M.A., Khatibi, R., Goel, A., FazeliFard, M.H., and Azani, A. 2016. Modeling river discharge time series using support vector machine and artificial neural networks. Environ Earth Sci.

Jajarmizadeh, M., Kakaei Lafdani, E., Harun, S., and Ahmadi, A. 2015. Application of SVM and SWAT models for monthly streamflow prediction, a case study in South of Iran. KSCE Journal of Civil Engineering, 19(1), 345-357.

Kecman, V. 2000. Learning and Soft Computing, Support Vector Machines, Neural Network and Fuzzy Logic Models. MIT Press, 2000 608p).

Khatibi, R., Naghipour, L., Ghorbani, M.A., and Aalami, M.T. 2012. Predictability of relative humidity by two artificial intelligence techniques using noisy data from two Californian gauging stations. Neural computing and application, pp. 643-941.

Misra, D., Oommen, T., Agarwal, A., Mishra, S.K., and Thompson, A.M. 2009. Application and analysis of support vector machine based simulation for runoff and sediment yield, Biosystems engineering, 103(2), 527–535.

Shin, S., Kyung, D., Lee, S., Taik Kim, J., and Hyun, J. 2005. An application of support vector machines in bankruptcy prediction model. Expert Systems with Applications, 28, 127-135.

Singh, V.P., Krstanovic, P.F., Lane, L.J., 1998. Stochastic models of sediment yield. In: Anderson, M.G. (Ed.), Modeling Geomorphological Systems, Vol. 2. John Wiley and Sons Ltd., 272–286.

Vapnik, V.N. 1998. Statistical Learning Theory. Wiley, New York.

White S. 2005. Sediment yield prediction and modeling. Hydrological Processes 19, pp.3053–3057.

Xu, L., Wang, J., Guan, J., and Huang, F. 2007. A Support Vector Machine Model for Mapping of Lake Water Quality from Remote-Sensed Images. IC-MED. 1(1), 57-66.

Yang, C.T. 1996. Sediment Transport, Theory and Practice. McGraw-Hill, New York.