Vol-4, Issue-2, Mar-Apr, 2020
ISSN: 2456-8635

Comparison and Evaluation of Support Vector Machine and Gene Programming in River Suspended Sediment Estimation (Case Study: Kashkan River)

Hamidreza Babbali[1*], Reza Dehghani[2]

[1]Assistant Professor of Civil Engineering, Islamic Azad University, Khorramabad, Iran
[2]Ph.D. Student of Water Structure, Faculty of Agric., University of Lorestan, Khorramabad, Iran

Abstract— *Simulation and evaluation of sediment are important issues in water resources management. Common methods for measuring sediment concentration are generally time consuming and costly and sometimes does not have enough accuracy. In this research, we have tried to evaluate sediment amounts, using Support Vector Machine (SVM), for Kashkanriver, Iran, and compare it with common Gene-Expression Programming. The parameter of flow discharge for input in different time lags and the parameter of sediment for output dhuring contour time (1998-2018) considered. Criteria of correlation coefficient, root mean square error, mean absolute error and Nash Sutcliff coefficient were used to evaluate and compare the performance of models. The results showed that two models estimate sediment discharge with acceptable accuracy, but in terms of accuracy, the support vector machine model had the highest correlation coefficient (0.994), minimum root mean square error (0.001ton/day) , mean absolute error(0.001 ton/day) and the Nash Sutcliff (0.988) hence was chosen the prior in the verification stage. Finally, the results showed that the support vector machine has great capability in estimating minimum and maximum sediment discharge values.*

Keywords— ***Suspended Sediment, Kashkan, Support Vector Machine, Gene Expression Programing.***

I. INTRODUCTION

Historically, there have been a number of attempts to estimate the sediment yield using modeling that can be broken down into different groups (White, 2005). The deterministic models can be grouped as 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, and there are some of these models 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 process computer 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 the flow Q and the 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 assume 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) 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 highly important aspect of neural computing because it allows it to be used to model a function where one has little information or incomplete understanding. Thus, the SVM approach is extensively used in the water resources literature in the field of prediction and forecasting (In recent years, Support Vector Machines (SVM) has been widely used in various fields. Runoff and sediment yield estimation can utilize SVM as well (misra*et al*,2009). SVM is a powerful nonlinear pattern recognition technique (Vapnik,1998; Kecman,2000). The relationship was used to estimate suspended sediment load by using linear regression model, power regression model, artificial neural network and support vector machine in this study. Records of river discharges and suspended sediment loads in Kaoping river basin were investigated as case study. The result shows that SVM outperforms the ANN and other two regression models(Chiang and Tsai,2011).This study presents 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. A functional relation has been developed using GEP. The proposed relationship can be applied to different boundaries with partial flow. The proposed GEP approach gives satisfactory results) compared to the existing predictor(Ghani and Azamathulla,2011).The study Records of river discharges and suspended sediment loads in the Goodwin Creek Experimental Watershed in United States were investigated as a case study. As a result, we believe that the proposed SVM model has high potential for predicting suspended sediment load(Chiang *et al*.,2014).The study compares the results of the Soil and Water Assessment Tool (SWAT) with a Support Vector Machine (SVM) to predict the monthly streamflow of arid regions located in the southern part of Iran, namely the Roodan watershed. Results indicate that the SVM has 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 includes different performance measures, which indicate that SVM and ANN have an edge over the results by the conventional RC and MLR models. Notably, peak values predicted by SVM and ANN are more reliable than those by RC and MLR, although the performances of these conventional models are acceptable for a range of practical problems(Ghorbani*et al*.,2016).In total, according to done research and mention this point that Kashkan river as the main source of water supply for different sectors and adjacent areas, so the estimated Suspended sediment and management proceedings to improved optimal operation of reservoir more than ever it is essential. So the purpose this research is estimated Suspended sediment in Kashkanriver with the help support vector machine and compared that's results with gene expression programming.

II. MATERIALS AND METHODS

Case study and used data

Kashkan River is the most flooded river in Lorestan province. The Kashkan catchment area is located in the southwestern part of Iran with a surface area of 1.5 km2. This area forms an important part of the rugged branches of the Karkhe River and covers about one-third of the Lorestan soil. Watershed of Kashan River in the hydrological division of Iran is a part of the Persian Gulf catchment. The river is located between latitude ′ 34 ′ 31 ° 47 ° to ′ 12 48 12 ° 48 east and latitude ″ 45 ° 5 ° 33 to ″ 41 ° 44 ° 33 ° N in Lorestan province. The location of the study area is shown in Figure 1.

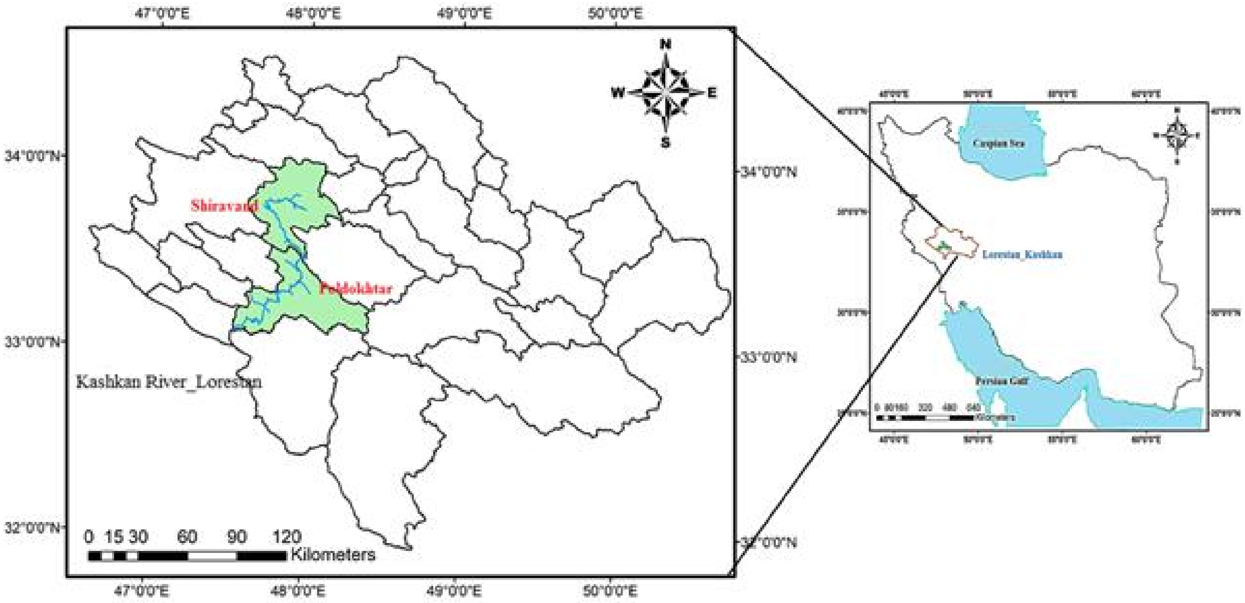


Fig.1: Geographical location Kashkanriver

Table 1. Statistical properties Sediment parameter month (1998-2018)

Station	Period of record	Data set	Statistics	Q(m3/s)	S(ton/day)
Kashkan	1998-2018	Training	Minimum	0.921	0.425
			Mean	18.865	4427.341
			Maximum	410.840	355474.284
			deviation	41.652	28324.672
			Skewness	7.542	12.424
		Testing	Minimum	0.000	0.742
			Mean	9.154	31.756
			Maximum	24.158	215.122
			deviation	7.874	52.105
			Skewness	1.058	2.598

Gene Expression Programming

Gene Expression Programming method presented with Ferreira in 1999 (Ferreira.2001). This method is a combination of genetic algorithms (GA) and genetic programming (GP) method than in this, simple linear chromosomes of fixed length are similar to what is used in genetic algorithm and branched structures with different sizes and shapes aresimilar to the decomposition of trees in genetic programming.Since this method all branch structures of different shapes and size are encoded in linear chromosome with fixed length, this is equivalent than Phenotype and Genotype are separated from each other and system could use all evolutionary advantagesbecause of their. Now,however the Phenotype in GEP included branch structures used in GP, but the branch structures be inferences by GEP (than also calledtreestatement) are explainer all independent genomes. In short can say improvements happened in linear structure then is expressed similar with tree structure and this causes only the modified genomemoved to the Next Generation and don't need with heavy structure to reproduce and mutation (Ferreira.2001). In this method different phenomena are modeling by collection of functions and terminals. Collection of functions generally include the main functions of arithmetic {+, -, ×, /}, the trigonometric functions or any other mathematical function {√, x2, sin, cos, log, exp, …} or defined functions by author whom believed they are appropriate for interpreting model. Collection of terminals consist problem's constants values and independent variables (2001). For applying gene expression programming method is used GenXproTools 4.0 Software. In order to obtain more information can recourse to (Ghorbani*et al*.,2012).

Support Vector Machine

Support Vector Machine is anefficient learning system based on optimization theory that used the principle of induction minimization Structural error and results an overall optimal solution(Vapnik,1998). In regression model SVM is estimated function associated with the dependent variable Y as if is afunction of several independent variables X(Xu*et al*.,2007).Like other regression problems is assumed the relationship between the dependent and independent variables to be determined with algebraic function similar f(x) plus some allowable error (ε).

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

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

If W is coefficients vector, b is constant characteristic of regression function, and also ∅ is kernel function, then goal is to find a functional form for f(x). It is realized with SVM model training by collection of samples (train collection). To calculate w and b require to be optimized error function in ε-SVM with considering the conditions embodied in Equation 4(Shin *et al*.,2005).

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

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

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

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

Average Lagrange Coefficients $\bar{\alpha}_i$ in characterized space is ∅(x).Maybe calculation be very complex. To solve this problem, the usual process of SVM model is choose a kernel function as followrelation.

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

Can be used of different kernel functions 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 respectively, are calculated as follows relation(Vapnik.1998).

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

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

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

Evaluation Criteria

In this research to evaluate the accuracy and efficiency of the models was 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 criterions are respectively 1, 0, 1, and 0.

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

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

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

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.

III. RESULTS AND DISCUSSION

One of the most important steps in modeling, is select the right combination of input variables. Also shown in Table 2.The structure of input combinations.

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 t time 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 sediment to Kashkan river use of different combinations of input parameters that showed them in Table3. To estimate input discharge Kashkan river using by Gene Expression Programming and Support Vector Machine with have catchment hydrometric data from 240 registered records during the period (1998-2018), count in 192 records to training and 48 remaining records to verification.

The results of Gene Expression Programming

Using gene expression programming due to the selection of variables in the model and remove variables with less impact and also ability to provide a clear relationship were considered to estimating sediment to the Kashkanriver. Since ever four input are incorporated to determining the significant variables and more reviews in addition three of the original operator (F1) and the states based on arithmetic operators default (F2). The reason for choice this type of operator has been based on studies (Ghorbani*et al.*,2012) and (Khatibi*et al.*,2012).

$$\mathrm{F1}:\{+,-,*,/,\sqrt{},\mathrm{Exp},\mathrm{Ln},{}^{2},{}^{3},\sqrt[3]{},\mathrm{Sin},\mathrm{Cos},\mathrm{Atan}\} \quad (13)$$

$$\mathrm{F2}:\{+,-,*,/\} \quad (14)$$

Results of gene expression programming model for both operator in Table3 show that F2 operator in both stages training and verification with maximum correlation coefficient R=0.813, root mean square error RMSE=0.002, mean absolute error MAE=0.002 and NS=0.643 has high accurate than other operators. Therefore gene expression programming with F2 operator include four the main mathematical operators with a simple mathematical relationship has the most accurate to estimating sediment to the Kashkan river.

Also, In order to compare the results of the usemodel support vector machine. The program for SVM was constructed using MATLAB (The MathWorksInc 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. In this study, the parameter search scheme employed is the shuffled complex evolution algorithm (SCE-UA), (see 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 to the Kashkanriver by SVM model can examine types of kernel function, than was selected linear kernel, polynomial and radial basis functions that are common types used in hydrology. The results of study models is given in Table3. According to this table combined model number 3 with radial basis functions kernel has the highest correlation coefficient R=0.994, lowest root mean square error RMSE=0.001 ton/day ,mean absolute errorMAE=0.001ton/day and NS=0.988 in verification stage that has optimal solution than other models. In Fig3 shown the best model for verification of data.

Table 3. The final results of the training and verification 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

As shown in Figure 2, scatter plot Support Vector Machine matched observed and simulated values relative to the with the best fit line there y=x. Which explains the ability of this model is the estimation most values. The scatter plots of gene expression programming related to the verification stage in Fig(2-b) show the fit line of computational values with four mathematical operators to the best fit line y=x.As is from this Fig, all of the estimated and observation values are in the fit line except few points that are not in bisector line which it is denoted the estimated and observed values of equality on the line (y=x). On the other side in figures 3, the graph of can be seen over time for simulation models. In fig 3-a , is shown Support Vector Machine model from performance acceptable in estimation values. But according to the fig 3-b, shown GEP model estimating the maximum acceptable accuracy has not been.

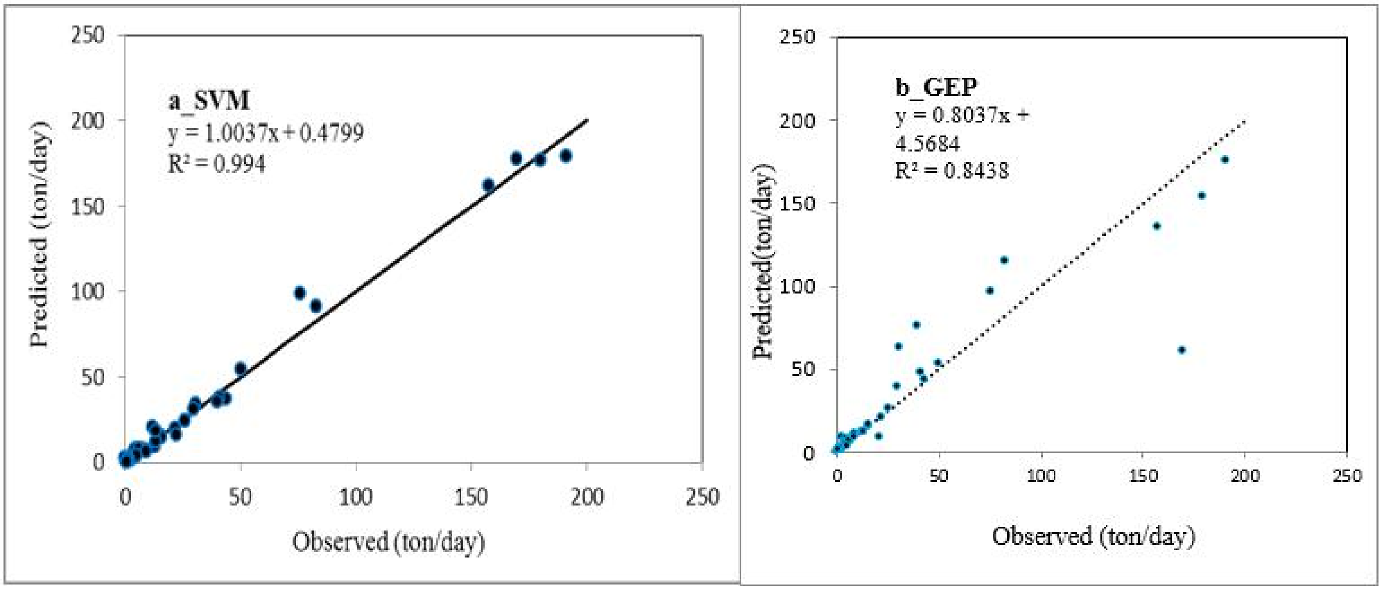
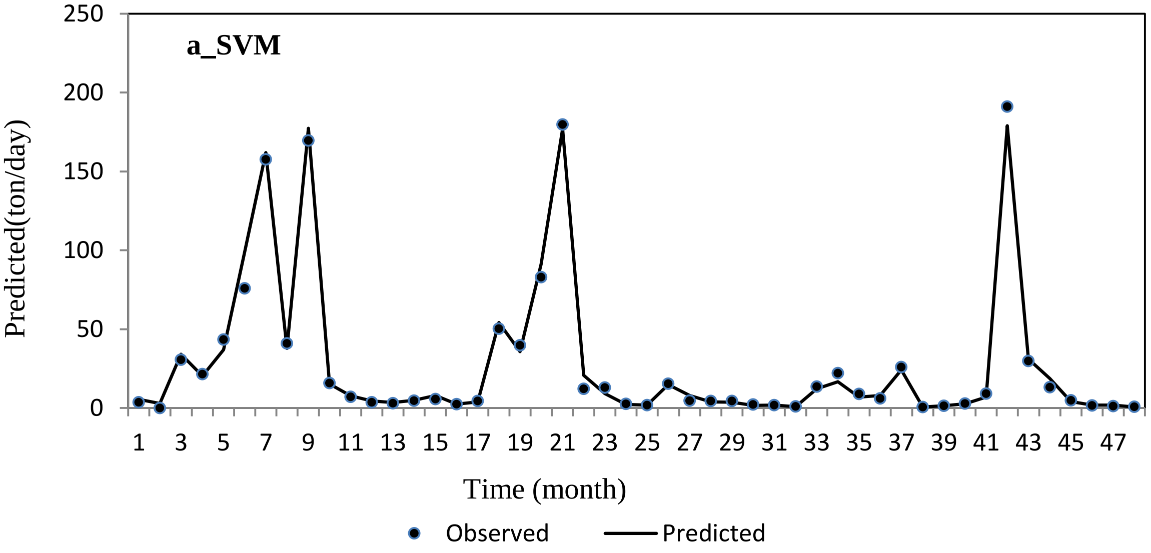


Fig.2: Scatterplots of the predicted-observed sediment time series of the Kashkan station in test period using (a)SVM; (b) GEP.



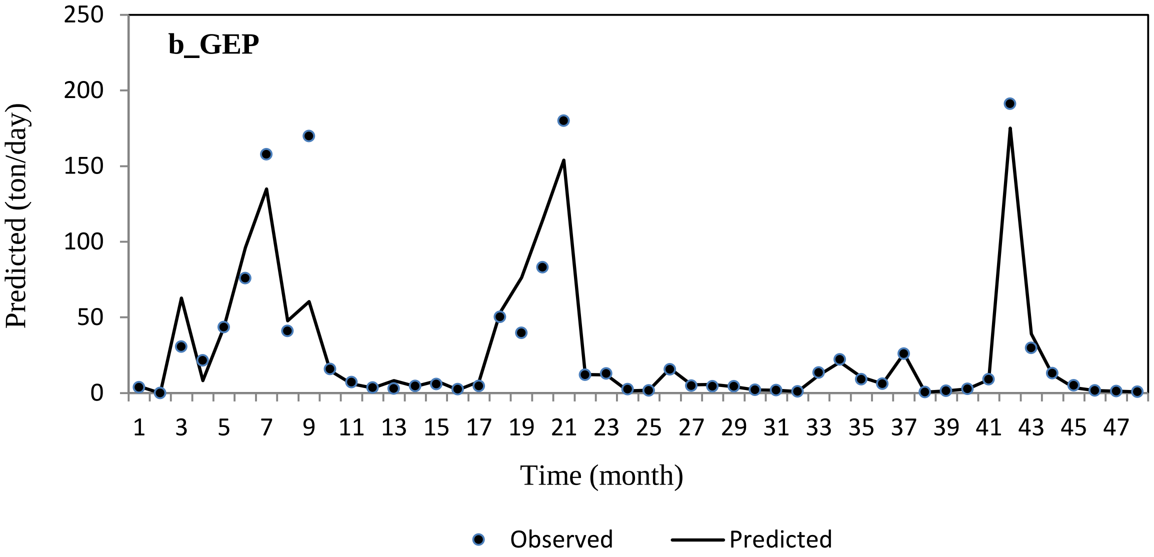


Fig.3: Comparison of the optimal models with observed values observed values plots for testing data set (a) SVM; (b) GEP

Comparison Performance of models

Choosing the optimal solution for each of the models and compare together was defined all two methods can with good accurate simulate sediment to the Kashkan river. Comparison of gene expression programming model and support vector machine model shown proximity the results of these two models. In Fig4 shown the results of all two models to the observed valueduring the time that all two models good function, whereas support vector machine modelis well covered minimum, maximum, and middle values.

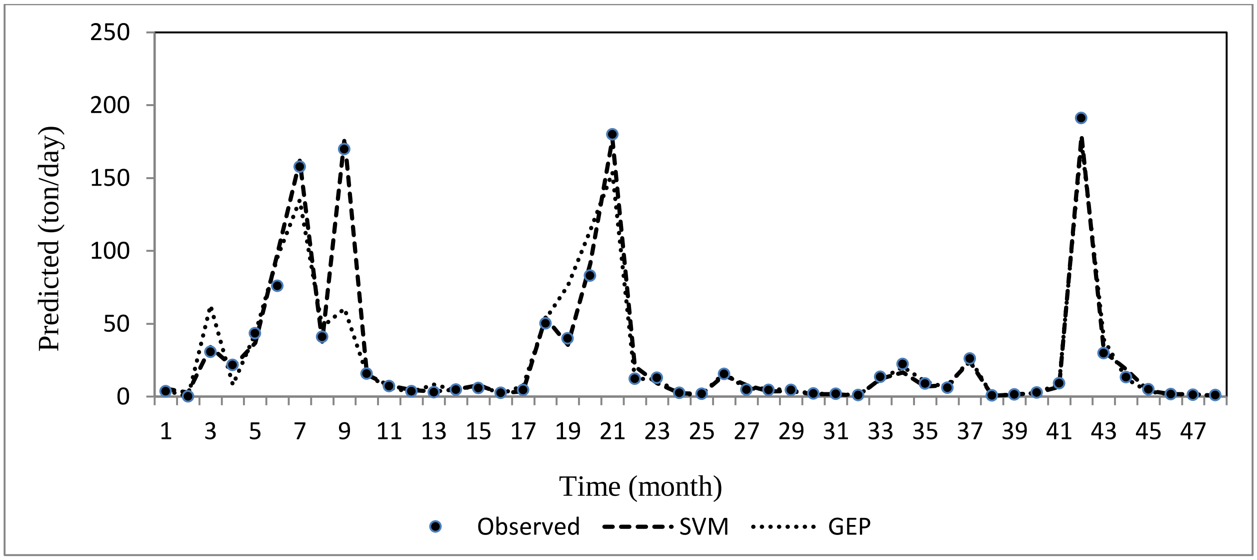


Fig.4: The scatter plot between estimated and observed values gene expression programming and support vector machine models for recorded data in verification stage

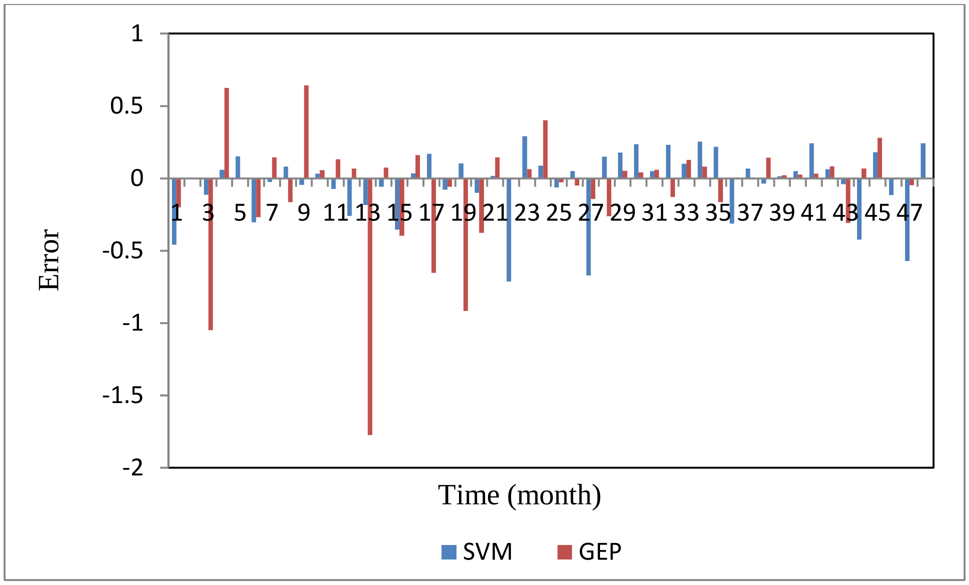


Fig.5: All two models graph optimization error as a percentage of the mean observed value

Finally difference between the observed sediment values and optimal computational models calculated as a percentage of the mean observed values (error value) and was drawn this diagram in comparison with the data recorded (Fig5). As seen in this Fig, more errors to ever two models has been ±5 band the highest error rate gene expression programming and support vector machine models are respectively 0.643 and 0.291 percent of the mean observed values. Among these models (GEP and SVM) svm model has lowest error value. Totally due to the high estimation accuracy and reliability gene expression programming and support vector machine models the correlation between the observed values and the computed values are respectively 0.994 and 0.813. Also the results of was significant estimated and observed values in the probability levels %5 and %10 shown, SVM model has significant correlation in both probability levels.

IV. CONCLUSIONS

In this research, we tried to evaluated performance some models to simulating sediment to the Kashkan River In the province lorestan using by sediment month data in Kashkanriver. Used models include gene expression programming and support vector machine models. Observed sediment values compared with estimated sediment in these models (GEP and SVM). The results summarized as follows:

A: SVM model has high accurate and a little error to estimate minimum, maximum, middle values and peak sediment, and high correlation with the observed value. B: Gene expression programming model with the four basic arithmetic operations has high ability to estimating minimum, maximum, and middle values and peak values, also support vector machine with radial basis functions kernel has high ability estimating minimum and middle values but to estimating maximum values do have enough operation. C: Increasing the number of parameters in the various models to simulating sediment cause to improve operation to estimating sediment. D: Estimating sediment using by combined models have lower error and high correlation than other models to estimated sediment in reservoirs dam.

Totally the results of this research showed support vector machine method has highest accurate than other models. As research results Jajarmizadeh *et al*(2015) and Ghorbani *et al*(2016) has been proven its. Also this research shown using of gene expression programming and support vector machine models could use to estimating sediment to the river.

V. ACKNOWLEDGMENTS

The authors are very grateful of the Regional Water Company, Lorestan Province, Iran, because they participate in gathering the data required to perform the work.

REFERENCES

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

[2] Chiang, J., Tsai, Y., 2011. Suspended Sediment Load Estimate Using Support Vector Machines in Kaoping River Basin. International Conference on suspended sediment load

[3] Cohn, T.A.,Caulder, D.L., Gilroy, E.J., Zynjuk, L.D., 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.pp.2353–2363.

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

[5] Forman, S.L., Pierson, J., 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,pp. 157– 176.

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

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

[8] Ghorbani, M.A., Khatibi, R., Asadi, H., Yousefi,P ., 2012. Inter- Comparison of an Evolutionary Programming Model of Suspended Sediment Time-series whit other Local Model. INTECH. doi. org/10.5772/47801,pp. 255-282

[9] Ghorbani, M.A., Khatibi, R.,Goel, A., FazeliFard, M.H., Azani, A., 2016. Modeling river discharge time series using support vector machine and artificial neural networks. Environ Earth Sci. DOI: 10.1007/s12665-016-5435-6

[10] Jajarmizadeh, M., KakaeiLafdani, E., Harun, S., 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):pp.345-357

[11] Kecman, V., 2000. Learning and Soft Computing, Support Vector Machines, Neural Network and Fuzzy Logic Models.MIT Press,2000(ISBN 0- 262-11255-8. 608p).

[12] Khatibi, R., Naghipour, L., Ghorbani, M.A., 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.

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

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

[15] 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., pp.272–286.

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

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

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

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