JRHS
Journal of Research in Health Sciences
journal homepage: www.umsha.ac.ir/jrhs

Original Article

Prediction the Groundwater Level of Hamadan-Bahar Plain, West of Iran Using Support Vector Machines

Lily Tapak (MSc)[a♠], Alireza Rahmani (PhD)[b♠] and Abbas Moghimbeigi (PhD)[c*]

[a] *Department of Biostatistics and Epidemiology, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran*

[b] *Department of Environmental Health Engineering, School of Public Health and Research Center for Health Sciences, Hamadan University of Medical Sciences, Hamadan, Iran*

[c] *Modeling of Noncommunicable Disease Research Center, Department of Biostatistics and Epidemiology, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran*

[♠] *The first and the second authors have same contribution.*

ARTICLE INFORMATION

Article history:

Received: *27 July 2013*

Revised: *08 September 2013*

Accepted: *24 September 2013*

Available online: *27 October 2013*

Keywords:

Groundwater Level

Support Vector Machine

Time Series

Iran

**** Correspondence***

Abbas Moghimbeigi (PhD)

Tel: *+98 811 8380398*

Fax: *+98 811 8380509*

E-mail1: *moghimbeigi@umsha.ac.ir*

E-mail2: *moghimb@yahoo.com*

ABSTRACT

Background: Water is considered as the main source of life but water resources are limited and nonrenewable. Different factors have caused groundwater to decrease. Therefore, modeling and predicting groundwater level is of great importance.

Methods: Monthly groundwater level data of about 20 years (October 1991 to February 2012) from the Hamadan-Bahar Plain, west of Iran were used based on peizometric height related to hydrologic years. The support vector machine (SVM), a new nonlinear regression technique, was used to predict groundwater level. The performance of the SVM model was assessed by using criteria of R^2, root mean square error (RMSE), means absolute error (MAE), means absolute percentage error (MAPE), correlation coefficient and efficiency coefficient (E) and was then compared with the classic time series model.

Results: The SVM model had greater R^2 (=0.933), E (=0.950) and Correlation (=0.965). Moreover, SVM had lower RMSE (=0.120), MAPE (=0.140) and MAE (=0.124). There was no significant difference between the estimated values using two models and the observed value.

Conclusions: The SVM outperforms classic time series model in predicting groundwater level. Therefore using the SVM model is reasonable for modeling and predicting fluctuations of groundwater level in Hamadan-Bahar Plain.

Citation: Tapak L, Rahmani AR, Moghimbeigi A. Prediction the Groundwater Level of Hamadan-Bahar Plain, West of Iran Using Support Vector Machines. J Res Health Sci. 2014;14(1): 82-87.

Introduction

Water is the main source of life while its resources are limited and nonrenewable. Groundwater is an important water resource for domestic, industrial, and agricultural activities [1]. During the last decades, due to the increasing demand for water, weak irrigation management and soil damage, the renewable groundwater level have decreased making a lot of worry for all. Regarding the limitation of water reservoirs, it is important to manage the use of water favorably for sustainable development [2]. Therefore, in order to effective management of groundwater, it is important to model and predict the fluctuations in groundwater level so that the authorities can plan prerequisites for reduction of unforeseeable variations.

Among the models used for this purpose, time series methods are usually applied to predict the groundwater level [3]. These methods have some limitations such as temporary prediction of variations. In recent years, soft computing tools such as support vector machines (SVMs) have been applied to predict in various areas of science and technology including time-series forecasting specially in water resource problems[4-9]. The SVM is considered as a nonlinear method based on minimizing structural risk and its ability have been verified in a variety of fields. In addition, as this method minimizes the experimental error and complexity simultaneously, it can improve its generalization for prediction [1]. Most researches have confirmed the positive performance of SVM method in prediction in surface water problems such as lake water level [10], flood and stream flow [11-13]. However, the application of SVM to groundwater level fluctuations modeling is limited. Khalil [14] used SVM for modeling of groundwater quality. Asefa[15] used SVM for developing long-term groundwater head monitoring networks. Yoon et al. [1] compared the neural network and SVM methods to predict groundwater level. Moreover, Shiri et al. [16] investigated the abilities of different data mining methods including SVM techniques for groundwater level forecasting and Sudheer et al. [17] proposed a SVM based method for groundwater level forecasting.

This study aimed to evaluate performance of SVM for groundwater level prediction in Hamadan-Bahar Plain, west of Iran and its results were compared with classic time series models.

Methods

The area of study

Hamadan-Bahar Plain, located in Hamadan, west of Iran with an approximate area of 930 km^2 at the sea level of 1700–1800 has spread over a vast valley surrounded by tall mountains. It has the least area among the plains of the region. Based on the climate diagram of Ambergheh, it is in the semiarid climate with a cold and mountainous weather. Regarding the increasing trend of consuming groundwater, population growth, undue use of water in agriculture, precipitation reeducation and land exploitation change along with soil permeability reduction, the level of groundwater has declined and annual feeding through precipitation has not been able to compensate the loss. In this study, monthly groundwater level longitudinal data of about 20 years (October 1991 to February 2012) from the Hamadan-Bahar Plain were used based on piezometric height related to hydrologic years provided through the Hamadan Water Corporation Regional Office.

ARIMA model

ARIMA model is a widely used model in the analysis of time series data. Autoregressive model (AR) provides a relationship between the current value of the variable Y_t and its past values Y_{t-k}[18]. Moving average model (MA) sets a relationship between the current value of variable and residuals in the past C_{t-k}, where k=1, 2,…. The ARIMA model is a mixture of AR and MA.ARIMA model can be formulated as

$$\phi_p(A)(1-A)^d Z_t = \theta_q(A)c_t$$

where p and q are the order of AR and AM models and d is the order of difference. In addition, we have:

$$\phi_p(A) = (1-\phi_1 A - \phi_2 A^2 - \ldots - \phi_p A^p)$$
$$\theta_q(A) = (1-\theta_1 A - \theta_2 A^2 - \ldots - \theta_q A^q).$$

To estimate the parameters of ARIMA, some methods are used such as maximum likelihood or conditional least square [19]. If two assumptions of white noise and normal distribution hold in ARIMA, this model can be place in good class [20]. The model calculations were done by SPSS16.

Support vector machine

The SVM was introduced by Vapnik[20,21] for regression and classification problems, based on the theory of statistical learning. It is known as support vector regression (SVR) in prediction problems. In this case we seek to find hyper plane close to most points. SVR differs from ordinary regression methods since it uses structural risk minimization instead minimizing empirical risk used in other learning theory methods like neural networks. It is therefore expected that this method outperforms other regression methods and is able to have better generalization [22].

Suppose we have $\{(x_i, d_i) : i = 1 \ldots n\}$, where x_i and d_i are input vector and desirable output respectively. The general form of estimation function of SVR is as $f(x) = (w.\psi(x)) + b$ in which w and b is coefficients and $\psi(x)$ is a nonlinear function in the feature space [22,23].

We can consider the nonlinear regression problem as an optimization problem in which we seek to minimize function[24] $R(C) = \frac{1}{2}\|w\|^2 + C\frac{1}{n}\sum_{i=1}^{n} L_\varepsilon(d_i, y_i)$ where

$$L_\varepsilon(d,y) = \begin{cases} |d-y|-\varepsilon & |d-y| \ge \varepsilon \\ 0 & o.w \end{cases}$$

The risk function

$$R(w,\xi^*) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)$$

can minimized in terms conditions

$$d_i - w\psi(x_i) - b_i \le \varepsilon + \xi_i$$
$$(w.\psi(x)) + b - d_i \le \varepsilon + \xi_i^*$$

in which $\xi_i, \xi_i^* \ge \circ$ are slack variables. After differentiation and some algebraic operations, the dual form of SVR can be rewritten as

$$\varphi(\beta_i, \beta_i^*) = \sum_{i=1}^{n} d_i(\beta_i - \beta_i^*) - \varepsilon\sum_{i=1}^{n} d_i(\beta_i + \beta_i^*) - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\beta_i - \beta_i^*)(\beta_j - \beta_j^*)K(X_i, X_j)$$

with the condition $\sum_{i=1}^{n}(\beta_i - \beta_i^*) = 0$, where $0 \le \beta_i, \beta_i^* \le C$ are Lagrange multiplier and $K(X_i, X_j)$ is kernel function. The kernel function decreases the dimension of calculation space. In SVMs, one of the functions polynomial, Radial basis Function, Linear and Sigmoid is considered to be kernel [24]. Similarly, parameters C and ε determined by user control the smoothness of approximated function and determine the error fluctuation borders, respectively [22]. The model calculations were done in MATLAB by LIBSVM library [27] and graphs were plotted using Microsoft Excel 2007.

Performance measures

Root mean square error (RMSE), mean absolute percentage error (MAPE), efficiency coefficient (E), determination coefficient (R^2), mean absolute error (MAE) and correlation coefficient (Corr) were used to assess the accuracy of SVM and ARIMA model. The R^2 and Corr measure the relationship between two variables. RMSE, MAE and MAPE provide different capabilities of prediction and goodness of fit. The efficiency coefficient measures the difference between observations and predictions to observed changes. The value of 90% or higher shows the satisfactory performance and under 80% the weak performance [28]. RMSE, MAPE, E AND MAE criteria are calculated by following relationships

$$RMSE = \sqrt{\frac{1}{n}\sum\left(y_{observed} - y_{predicted}\right)^2}$$

$$MAPE = 100 \times \frac{1}{n}\sum\frac{\left|y_{observed} - y_{predicted}\right|}{y_{observed}}$$

$$MAE = \frac{1}{n}\sum\left|y_{observed} - y_{predicted}\right|$$

$$E(\%) = 1 - \frac{\sum\left(y_{observed} - y_{predicted}\right)^2}{\sum\left(y_{observed} - y_{mean}\right)^2}$$

Where n is the number of observation, is monthly groundwater level and y_{mean} is the average observed groundwater level.

Results

Monthly groundwater level in Hamadan-Bahar Plain were used in this study based on peizometric height from 1991 to 2012 related to hydrologic years. To assess and compare the model performance, cross validation was used in which the data set was divided into training (82%) and test (18%) sets.

Summary statistics of data set are presented in Table 1, which shows data of test and training set are relatively homogenous. Besides, Figure 1(a) shows the general trend of hydrograph, representing the descending reduction of water reservoirs.

The correlation analyses was used to choose appropriate time series model and the auto-correlation, the partial auto correlation statistics and related confidence band of 95% were estimated for groundwater level in Hamadan–Bahar plain (Figure 1(b)). The auto correlation function (PACF) shows a significant correlation to lag2 and then it lies in confidence limits approximately. The pattern of rapid reduction of PACF shows the superiority of autoregressive process versus the moving average process. Therefore, SAR model was used to analyze the data based on graphs and seasonal variations. Two hypothesis of uncorrelated errors and zero mean errors was evaluated. With regard to the zero mean condition of residuals, the value of the statistic was 0.016 (P>0.05), so the hypothesis of zero mean errors and based on Ljung-Box statistics [20], uncorrelated errors assumption was not rejected (P>0.05).

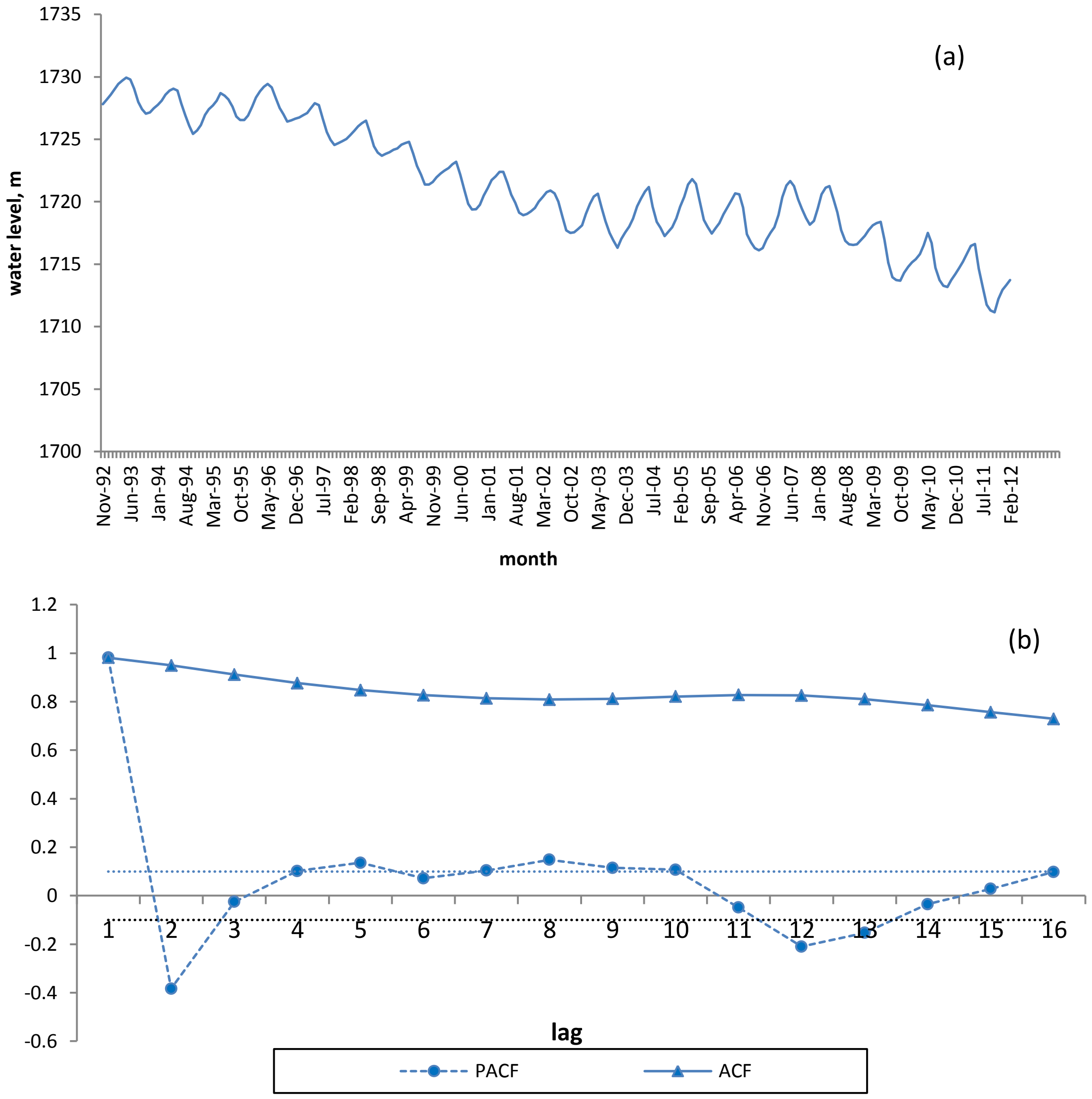

Figure 1: (a) Hydrograph of groundwater in Hamadan-Bahar Plain in hydrological years, (b) Auto-correlations function (ACF) and partial Auto-correlations function (PACF) of groundwater level

Table 1: Statistical parameters of groundwater level data

Water level	N	Min	Max	Mean	SD	Skewness
Total data	245	1711.15	1730.56	1721.70	4.96	0.16
Training set	200	1716.11	1730.56	1723.20	4.19	0.17
Test set	45	1711.15	1719.15	1718.20	4.01	0.20

The data having been normalized, SVM model was fitted to data by using RBF kernel function. The reason of its selection is based on the fact that significant difference for SVM using four functions of linear kernel function, polynomial sigmoid and RBF was not observed. Moreover, the process of finding best parameters to minimize the structural risk function was preformed through trial and error. The optimal values were obtained for three parameters c, ε, δ in RBF kernel function as 70, 0.01, and 0.25, respectively.

The correlation coefficient, R^2, RMSE, MAPE, E and MAE statistics of two models for train and test sets are given in Table 2. Results showed that the SVM model outperforms classic time series model (R^2=0.933 for SVM versus R^2=0.909 for SAR, RMSE=0.120 for SVM versus RMSE=.495 for SAR, E=0.950 for SVM versus E=0.910 for SAR, MAPE=0.140 for SVM versus MAPE=0.210 for SAR, MAE=0.124 for SVM versus MAE=0.490 for SAR and Correlation=0.965 for SVM versus Correlation=0.953 for SAR).

The variations of groundwater level along with estimated values from SVM model with RBF kernel function and classic time series for test data set are depicted in Figure 2. Both methods have proper performance, but the similarity of estimations of SVM to real values is greater and approximately overlaps. The residuals of two models are also drawn in this figure. It is clearly seen that the residuals from SVM model is less than those of classic time series, with less variations. Figure 3 shows the scatter plot of estimations from SVM (a) and classic time series (b) model versus observed values. Accordingly, the estimation values of SVM model have stronger relationship with real values. According to scatter plots, we can see that the estimations of classic time series have more scattered and farther than those of SVM model to real values.

Table 2: Statistics of R^2, root mean square error (RMSE), means absolute error (MAE), means absolute percentage error (MAPE), correlation coefficient (Corr) and efficiency coefficient (E) for support vector machine and classic time series performance in groundwater level

Method	R^2	RMSE	E	MAPE	MAE	Corr
SVM						
Train set	0.994	0.095	0.999	0.010	0.093	0.996
Test set	0.933	0.120	0.950	0.140	0.124	0.965
Classic time series						
Train set	0.953	0.194	0.985	0.034	0.360	0.976
Test set	0.909	0.495	0.910	0.210	0.490	0.953

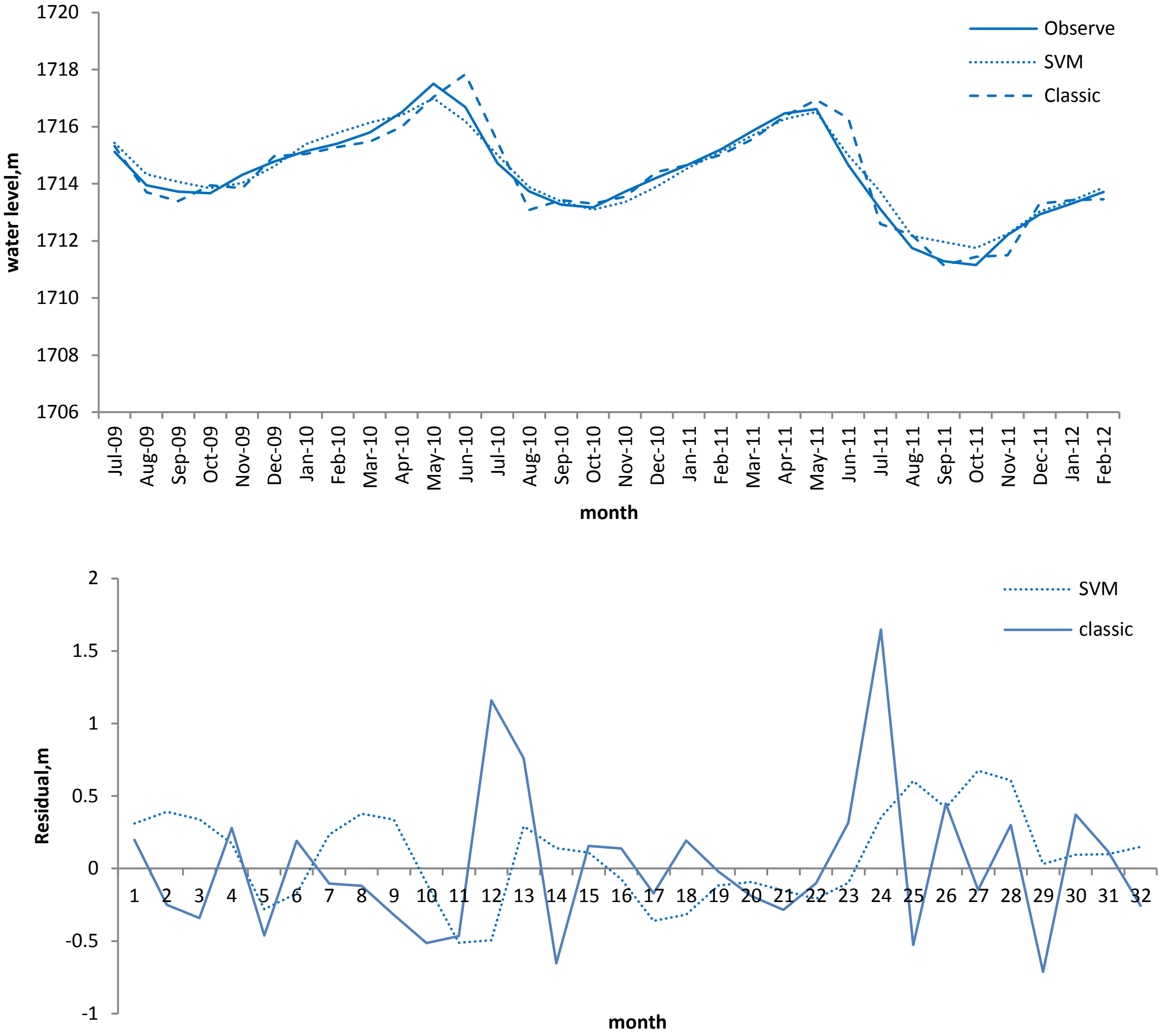

Figure 2: Monthly (a) estimations and (b) residuals of groundwater level based on Support Vector Machine (SVM) and classic time series

To test the significance of difference between the estimated values using two models and the observed values, paired *t*-test was carried out with significance level of 0.05. There was no significant difference between estimated values from SVM model (t_0=0.025, *P*=0.980) and classic time series (t_0=-0.128, *P*=0.898) and corresponding real values. The SVM model gives greater p-values. Based on the results of this test, the similarity between estimated values using SVM model and real values is more significant, thus, it is more powerful than classic time series in predicting groundwater level.

Discussion

The results of this study showed that SVM is a useful method for the empirical forecasting of groundwater level in Hamadan-Bahar Plain. The comparison between SVM and classic time series models according to different criteria showed that the SVM predictions are better.

In similar studies, superior performance of this method was verified compared with the neural network method to predict the groundwater level [1,17] and lake level fluctuations [24]. In a study conducted by Kandananond[27] performance of support vector machine, artificial neural network and ARIMA was compared to forecasting actual demand of different products from a consumer product company in Thailand and results indicated that the SVM method had a better forecast quality (in terms of MAPE) than ANN and ARIMA in every category of products which is similar to our results. Pai and Lin[28] report in their comparison study of ARIMA and the SVMs models (in terms of MAE, RMSE, MAPE and MSE) in forecasting stock prices problems better performance of SVM than ARIMA which is consistent with our results. Yu and Liong[29] in a study to predict Tryggevælde catchment runoff and Mississippi river flow showed that the SVM performance was better than ARIMA which is consistent with our results. Sapankevych and Sankar[30] evaluated different aspect of SVM for predicting time series data and showed powerful performance of SVM. Yoon et al.[1] conducted a comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer. Their result showed superiority of SVM than artificial neural networks. Likewise in similar study for comparison of artificial neural networks and support vector machine techniques in modeling lake level fluctuations in Turkey, Çimen and Kisi[26] showed better performance of SVM than artificial neural networks. In a study conducted by Rahmani et al. [2], a time series model based on 1984-2003 data set was used to predict groundwater level in this area and the water level was predicted to decline.

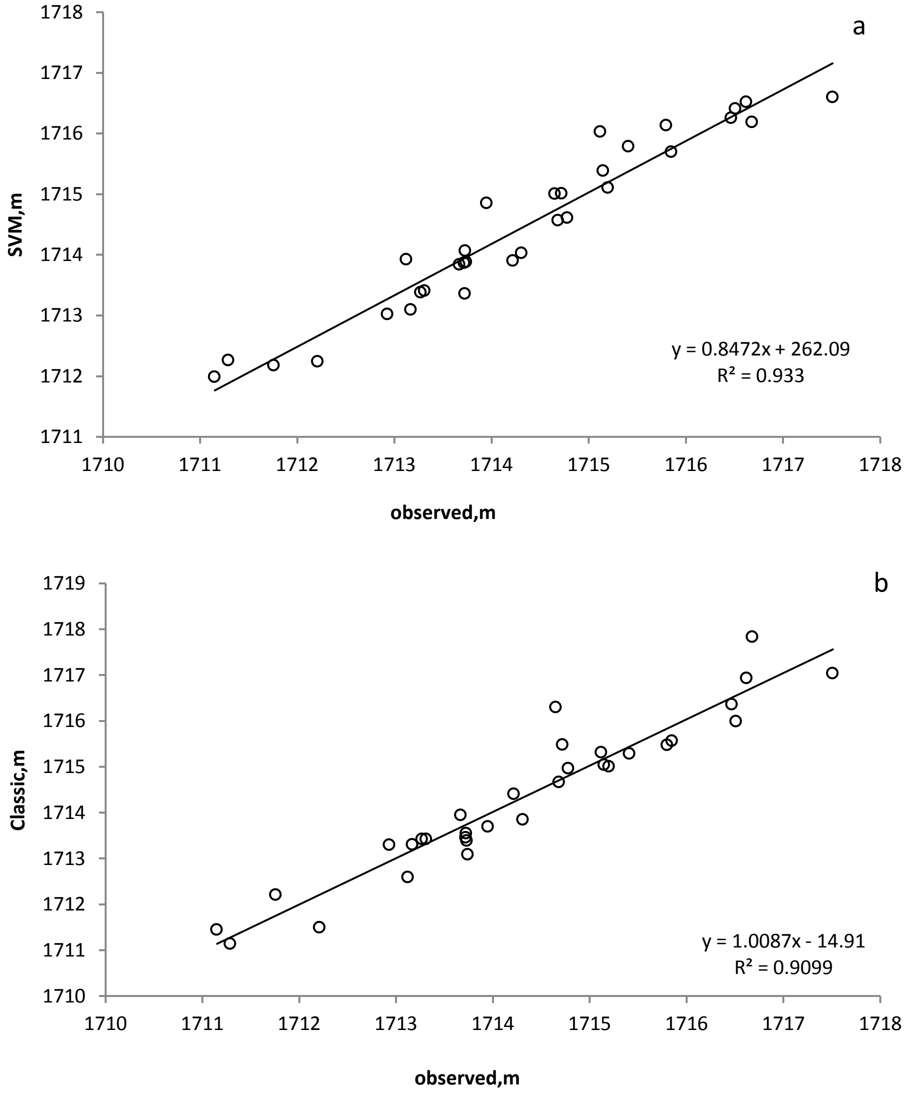

Figure 3: Comparison of groundwater level predictions from (a) support vector machine (SVM) and (b) classic time series models with observed values in Hamadan-Bahar Plain

Based on SVM model by using training set for predicting the test set, a decreasing trend was predicted for the plain groundwater level changes. This trend is expected to continue in the future and we are still faced with the loss of water.

According to the volatility seen in the graph and predicted seasonal trends, it can be inferred that the main source of variation is seasonal fluctuations like rainfall and taken much of the sources of lowland and any variation in these two could have a significant impact on the trend in the future. Therefore it can be reduced ascending trend with provision such as reducing per capita water or artificial recharge of aquifers.

Hence the overall strategies like control of external factors (mainly climatic characteristics, the topography, steep terrain and geological structure of the area) and human factors (overuse of water, especially in agriculture) affecting plain groundwater resources or artificial recharge of aquifers to revive is recommended to reduction of groundwater level in Hamadan-Bahar Plain.

Conclusions

This study focuses on the performance of SVM for groundwater level forecasting. By choosing appropriate kernel function and related parameters, SVMs can be a useful tool for nonlinear predicting problem especially with an unknown distribution. This study demonstrated that SVM can be used as an alternative of classic time series models for forecasting groundwater level. However, further research is required.

Acknowledgments

We like to thanks the Hamadan Water Corporation Regional Office for registering data.

Conflict of interest statement

The authors declare that there is no conflict of interests.

Funding

The Authors received no funding for this study

References

1. Yoon H, Jun SC, Hyun Y, Bae GO, Lee KK. A comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer. *J Hydrol.* 2011;396(1-2):128-138.

2. Rahmani A, Sedehi M. Prediction of groundwater level changes in the plain of hamedan-Bahar Using time series model. *J Water Wastewater Res.* 2004;15(51):42-49.

3. Tankersley CD, Graham WD, Hatfield K. Comparison of univariate and transfer function models of groundwater fluctuations. *Water Resour Res.* 1993;29(10):3517-3533.

4. Gao JB, Gunn SR, Harris CJ, Brown M. A probabilistic framework for SVM regression and error bar estimation. *Machine Learning.* 2002;46(1-3):71-89.

5. Hastie T, Tibshirani R, Friedman J, Franklin J. The elements of statistical learning: data mining, inference and prediction. *Math Intelligencer.* 2005;27(2):83-85.

6. Vapnik V, Golowich SE, Smola A. Support vector method for function approximation, regression estimation, and signal processing. *Adv Neural Inf Process Syst*. 1997;9(9):281-287.

7. Noori R, Karbassi AR, Moghaddamnia A, Han D, Zokaei Ashtiani MH, Farokhnia A, et al. Assessment of input variables determination on the SVM model performance using PCA, Gamma test, and forward selection techniques for monthly stream flow prediction. *J Hydrol.* 2011;401(3-4):177-189.

8. Tripathi S, Srinivas VV, Nanjundiah RS. Downscaling of precipitation for climate change scenarios: A support vector machine approach. *J Hydrol.* 2006;330(3-4):621-640.

9. Wu CL, Chau KW. Prediction of rainfall time series using modular soft computing methods. *Eng Appl Artif Intel.* 2013;26(3):997-1007.

10. Dibike YB, Velickov S, Solomatine D, Abbott MB. Model induction with support vector machines: introduction and applications. *J Computing Civil Eng.* 2001;15(3):208-216.

11. Asefa T, Kemblowski M, McKee M, Khalil A. Multi-time scale stream flow predictions: The support vector machines approach. *J Hydrol.* 2006;318(1):7-16.

12. Liong SY, Sivapragasam C. Flood stage forecasting with support vector machines1. *Jawra J Am Water Resour As.* 2002;38(1):173-186.

13. Yu PS, Chen ST, Chang IF. Support vector regression for real-time flood stage forecasting. *J Hydrol.* 2006;328(3):704-716.

14. Khalil AF, McKee M, Kemblowski M, Asefa T, Bastidas L. Multiobjective analysis of chaotic dynamic systems with sparse learning machines. *Adv in Water Resour.* 2006;29(1):72-88.

15. Asefa T, Kemblowski MW, Urroz G, McKee M, Khalil A. Support vectors–based groundwater head observation networks design. *Water Resour Res.* 2004;40(11):1-14.

16. Shiri J, Kisi O, Yoon H, Lee KK, Hossein Nazemi A. Predicting groundwater level fluctuations with meteorological effect implications—A comparative study among soft computing techniques. *Comput Geosci.* 2013;56:32-44.

17. Sudheer C, Shrivastava NA, Panigrahi BK, Mathur S. Groundwater level forecasting using SVM-QPSO. *Swarm, Evolutionary, and Memetic Computing*. 2011;7076:731-741.

18. Wei WWS. *Time series analysis*. 2nd ed. Boston: Pearson Addison Wesley; 2006.

19. Box GE, Jenkins GM, Reinsel GC. *Time series analysis: forecasting and control*: 4th ed. John Wiley & Sons; 2008.

20. Vapnik VN. *Statistical learning theory*. New York: Wiley; 1998.

21. Vapnik VN. *The nature of statistical learning theory*: 2nd ed. New York: Springer; 2000.

22. Radhika Y, Shashi M. Atmospheric temperature prediction using support vector machines. *J Comput Theory and Eng.* 2009;1(1):1793-8201.

23. Collobert R, Bengio S. SVMTorch: Support vector machines for large-scale regression problems. *J Mach Learn Res.* 2001;1:143-160.

24. Çimen M, Kisi O. Comparison of two different data-driven techniques in modeling lake level fluctuations in Turkey. *J Hydrol*. 2009;378(3):253-262.

25. Chang CC, Lin CJ: LIBSVM: a library for supportvector machines [updated March, 2013; cited October, 2013]; Avaliable from: http://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf.

26. Legates DR, McCabe GJ. Evaluating the use of “goodness-of-fit” measures in hydrologic and hydroclimatic model validation. *Water Resour Res.* 1999;35(1):233-241.

27. Kandananond K. A comparison of various forecasting methods for autocorrelated time series. Int. *J Eng Bus Manag.* 2012;4(4):1-6.

28. Pai PF, Lin CS. A hybrid ARIMA and support vector machines model in stock price forecasting. *Omega.* 2005;33(6):497-505.

29. Yu X, Liong SY. Forecasting of hydrologic time series with ridge regression in feature space. *J Hydrol.* 2007;332(3):290-302.

30. Sapankevych N, Sankar R. Time series prediction using support vector machines: a survey. *Ieee Comput Intell M.* 2009;4(2):24-38.