Journal of AI and Data Mining
Vol 5, No 2, 2017, 319-325

Drought Monitoring and Prediction using K-Nearest Neighbor Algorithm

E. Fadaei-Kermani*, G. A. Barani and M. Ghaeini-Hessaroeyeh

Department of Civil Engineering, Faculty of Engineering, Shahid Bahonar University of Kerman, Kerman, Iran.

Received 16 April 2016; Received 07 December 2016; Accepted 24 December 2016
**Corresponding author: E-mail: ehsanhard@gmail.com (E. Fadaei-Kermani).*

Abstract
Drought is a climate phenomenon that might occur in any climate condition and all regions on the earth. An effective drought management depends on the application of appropriate drought indices. Drought indices are variables that are used to detect and characterize drought conditions. In this work, it is tried to predict drought occurrence based on the standard precipitation index (SPI) using k-nearest neighbor modeling. The model is tested using the precipitation data of Kerman, Iran. The results obtained show that the model gives reasonable predictions of the drought situation in the region. Finally, the efficiency and precision of the model is quantified by some statistical coefficients. Appropriate values for the correlation coefficient (r = 0.874), mean absolute error (MAE = 0.106), root mean square error (RMSE = 0.119) and coefficient of residual mass (CRM = 0.0011) indicate that the presented model is suitable and efficient.

Keywords: *Drought Monitoring, Standard Precipitation Index, Nearest Neighbor Model, Model Evaluation.*

1. Introduction

Drought is a natural and repeatable phenomenon caused by a decline in the rainfall level during a specified time period. This phenomenon is a climatic event because its characteristics depend on its intensity and continuation as well as the extent of the affected area. Its occurrence can be short-term and harmless or harmful and long-term. It starts slowly, and its effects appear gradually and in a relatively long period of time in different sectors such as water resources, agriculture, environment and economy. Therefore, determining the exact starting and ending points of this phenomenon is rather difficult. That is why drought has been often described as a creeping phenomenon [1].

Drought monitoring and forecasting play a crucial role in the management of water resource systems, and can considerably reduce the losses caused by this phenomenon. Generally, drought indices are used to monitor and predict this phenomenon. The overall objective of these indices is to express this phenomenon quantitatively and to incorporate the combined effects of various factors on the occurrence of droughts in the form of more quantitative and convenient relationships [2].

A number of different indices have been developed to monitor and quantify a drought, each with its own characteristics. They include the Palmer drought severity index (PDSI; Palmer [3]), rainfall anomaly index (RAI; Van Rooy [4]), deciles (Gibbs and Maher [5]), crop moisture index (CMI; Palmer [3]), Bhalme and Mooly drought index (BMDI; Bhalme and Mooley [6]), surface water supply index (SWSI; Shafer and Dezman [7]), national rainfall index (NRI; Gommes and Petrassi [8]), standardized precipitation index (SPI; McKee et al. [9]), reclamation drought index (RDI; Weghorst [10]).

Examples of the drought damage to agricultural systems and other sectors around the world are well-documented, and various efforts have been made to investigate and characterize the mechanism of this phenomenon. Cancelliere et al. [11] have provided two methodologies for the seasonal forecasting of SPI, under the hypothesis of uncorrelated and normally distributed monthly precipitation aggregated at various time scales. Han et al. [12] have proposed a method for

drought forecasting based on the remote sensing data using the ARIMA models. The method was used for drought forecasting in the Guanzhong Plain. Farokhnia et al. [13] have utilized the adaptive neurofuzzy inference system (ANFIS) model to forecast possible drought conditions in Tehran plain. Du et al. [14] have defined the synthesized drought index (SDI) as a principal component of vegetation condition index (VCI), temperature condition index (TCI) and precipitation condition index (PCI) for drought monitoring in Shandong province, China. Farahmand and AghaKouchak [15] have introduced the Standardized Drought Analysis Toolbox (SDAT), which can be applied to different climatic variables including precipitation, soil moisture, and relative humidity without having to assume representative parametric distributions. Hao et al. [16] have proposed the optimized meteorological drought index (OMDI) and the optimized vegetation drought index (OVDI) using the multi-source satellite data to monitor drought in three bio-climate regions of SW China.

Non-parametric methods can be used as appropriate approaches to estimate the status of droughts. In the cases where the relationship between input and output is not already fully-determined, utilizing non-parametric algorithm can be instrumental. Therefore, in this study, using the nearest neighbor model, a method was applied to monitor and predict droughts based on the standard precipitation index.

2. Standard precipitation index (SPI)

The understanding that a deficit of precipitation can have different impacts on the ground water, reservoir storage, soil moisture, and streamflow led McKee et al. [9] to develop the Standardized Precipitation Index (SPI) to enhance the detection of onset and monitoring of drought for multiple time scales. These time scales reflect the impact of drought on the availability of the different water resources. Soil moisture conditions respond to precipitation anomalies on a relatively short scale, while ground water, streamflow, and reservoir storage reflect the longer term precipitation anomalies. The standardized precipitation index (SPI) was calculated, based on the long-term precipitation record for a desired period (at least 30 years). The long-term record was fitted to a probability distribution, most probably gamma distribution. Then the cumulative probability was transformed to a *Z*-standard normal distribution with mean zero and variance of one using the following equations [9,17]:

$$Z = SPI = -[t - \frac{C_0 + C_1 t + C_2 t^2}{1 + d_1 t + d_2 t^2 + d_3 t^3}] \quad (1)$$

$$t = \sqrt{\ln[\frac{1}{H(x)^2}]} \qquad 0 < H(x) \le 0.5$$

$$Z = SPI = +[t - \frac{C_0 + C_1 t + C_2 t^2}{1 + d_1 t + d_2 t^2 + d_3 t^3}] \quad (2)$$

$$t = \sqrt{\ln[\frac{1}{1 - H(x)^2}]} \qquad 0.5 < H(x) \le 1$$

where, $H_{(x)}$ is the cumulative probability function, and the constants C_1 to C_3 and d_1 to d_3 can be calculated as follows:

$C_1 = 2.51557$ $\quad d_1 = 1.432788$

$C_2 = 0.802853$ $\quad d_2 = 0.189269$

$C_3 = 0.010328$ $\quad d_3 = 0.001308$

Using the time series obtained by precipitation data, sorting data in increasing order, the empirical probability distribution can be calculated (Eq. 3).

$$ECP = \frac{m}{n+1} \quad (3)$$

where, *m* is the row number of sorted precipitation data and *n* is the total number of precipitation data. Using the standard normal cumulative distribution curve, the standard precipitation index (SPI) can be calculated related to the precipitation data for every corresponding time scale.

Table 1 shows the classification system defining drought intensities resulting from SPI. According to this table, drought occurs any time SPI is continuously negative and reaches intensity where SPI is -1.0 or less. The drought event ends when SPI becomes positive. Each drought event, therefore, has a duration defined by its beginning and end, and the intensity for each month that the event continues.

SPI has several advantages over other indices including its simplicity and temporal flexibility, which allow its application for water resources on all timescales. Moreover, as SPI is adaptable for the analysis of drought at variable time scales, it can be used for monitoring agricultural and hydrological aspects [18]. Despite all these advantages, this index has some limitations as well. SPI uses only the precipitation data, and it is loosely connected to ground conditions [19].

3. K-Nearest neighbor modeling

The *k*-nearest neighbor modeling (*k*-NN) is a nonparametric machine learning algorithm that has found wide usage in pattern recognition and

data mining. It tries to classify an unknown sample based on the known classification of its neighbors. In this method, the model is fed with a training set, and it uses this training set to classify objects. Each one of the samples in the training set is labeled. The input objects are classified based on the *K* parameter, meaning that they are assigned to the class that is most pervasive among its closest *K* neighbors [20,21].

Table 1. Drought classification of SPI [9].

Class	SPI Values	Drought Status
1	+2 and more	extremely wet
2	1.5 to 1.99	very wet
3	1 to 1.49	moderately wet
4	0.99 to -0.99	near normal
5	-1.49 to -1	moderately dry
6	-1.99 to -1.5	very dry
7	-2 and less	extremely dry

Despite its simplicity, the k-NN algorithm has been widely studied from various perspectives, pursuing the improvement of its classification accuracy. The *K*-NN modeling has been used for traffic flow forecasting [22], streamflow simulation [23, 24], prediction of intake vortex risk [25], and prediction of cavitation damage on dam spillways [26]. The advantages of the *K*-NN model are (1) it is highly effective, especially where use is made of large datasets; (2) algorithm efficiency allows various combinations of the factors to be tested, and insignificant combinations are detected and eliminated, minimizing the risk of overfitting; and (3) the *K*-NN algorithm is robust even where noisy data is used [20, 25].

The first step in the *K*-NN model is to find the distance between the training and test data. The choice of the distance measure is an important consideration. Commonly, the Euclidean distance measure is used (Eq. 4).

$$d(X,Y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \quad (4)$$

where X refers to the training data with specific parameters (x_1 to x_n), and Y refers to the test data with specific parameters (y_1 to y_n).

$$X=(x_1,x_2,\ldots,x_n)$$

$$Y=(y_1,y_2,\ldots,y_n)$$

The next step involves sorting the distances for all the objects in the training set and determining the nearest neighbor based on the minimum distance (maximum similarity). The most important step in this model is to identify the *K* parameter, which is the number of the closest neighbors in the space of interest. If *K* is too large, classes with a great number of classified samples can overwhelm small ones and the results will be biased or the neighborhood may include too many points from other classes. On the other hand, if *K* is too small, the advantage of using many samples in the training set is not exploited, and the result can be sensitive to noise points [27,28].

The best value for *K* can be obtained by the *n*-fold cross-validation method. In this method, the data set is divided into *K* roughly equal-sized parts. For the *K*th part, the model is fitted to the other *K*-1 parts of the data, and calculating the prediction error of the fitted model when predicting the *k*th part of the data. This is done for all values of *K* (k= 1, 2, …, *K*), combining the *K* estimates of prediction error [27,29].

4. Model preprocessing and methodology

In this work, the precipitation data of the city of Kerman during 1995 to 2005 was used. This city is the capital city of the Kerman Province, which is located in the SE of Iran, situated on a sandy plain with 1749 meters above the sea level, and has an area of 181,714 km².

Based on the precipitation data, determining the moving time series and standard normal distribution function for different time scales, the standard precipitation index was calculated. Figs. 1- 3 show the precipitation cumulative probability distribution function and standard normal probability distribution function for 3-, 6-, and 12-month time scales. Finally, the SPI values were calculated for different periods of 3-month, 6-month, and 12-month during 1995 to 2005. Figs. 4 and 5 show the SPI values for the first 3- and 6-month time scales.

Before working with the *K*-NN model, to avoid bias toward one attribute or the other, the data is required to be normalized. Therefore all the input attributes are transformed to obtain temporary variables with a distribution having zero means and a standard deviation of 1 using the following equation:

$$X' = \frac{x - \bar{x}}{\sigma(x)} \quad (5)$$

where, X' represents the value for the normalized attribute, and $\bar{x}$ and $\sigma(x)$ represent the mean and standard deviations of the observed value of the attribute in the reference data set, respectively.

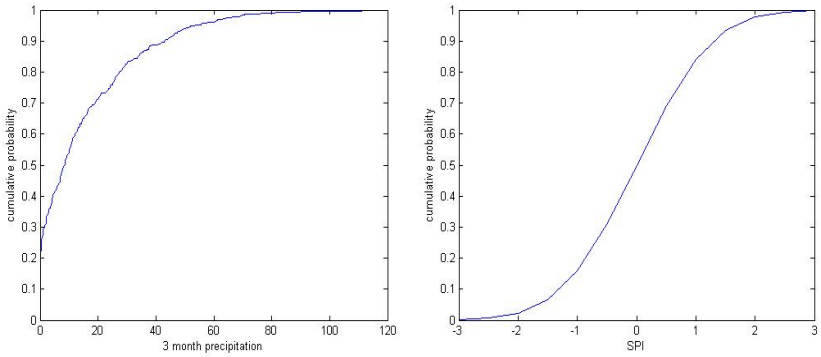

Figure 1. 3-month SPI values.

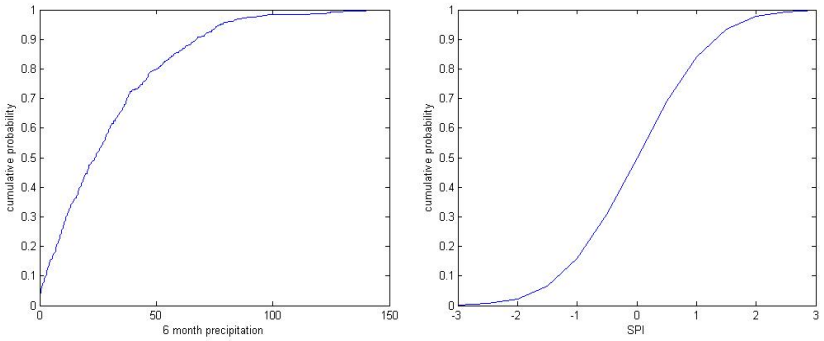

Figure 2. 6-month SPI values.

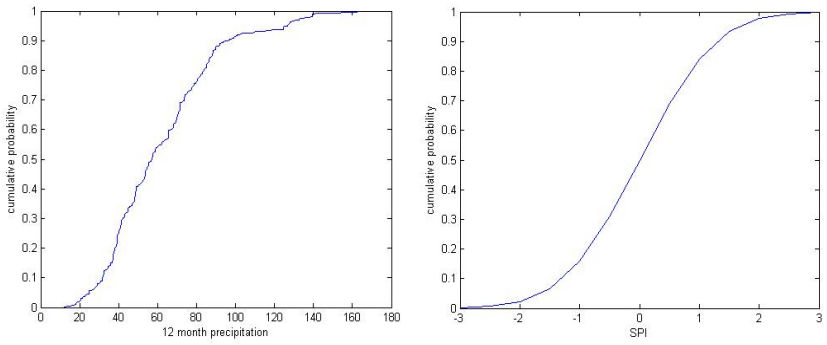

Figure 3. 12-month SPI values.

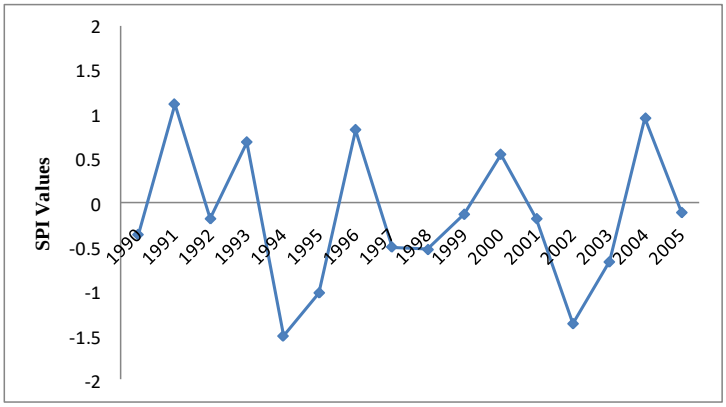

Figure 4. SPI values for first 3-month.

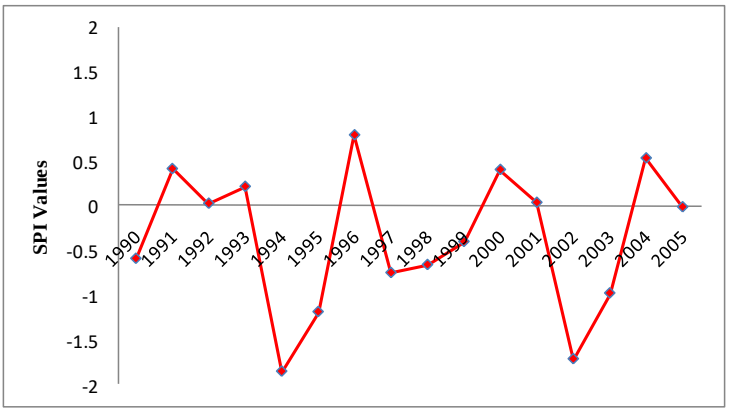

Figure 5. SPI values for first 6-month.

Finally, the efficiency and precision of the model could be evaluated by some statistical coefficients. The Pearson correlation coefficient (r) is a measure indicating the strength and direction of a linear relationship between two variables (model output and observed values). The Pearson correlation coefficient can be obtained by (6).

$$r = \frac{n[\sum_{i=1}^{n} y_i x_i] - [\sum_{i=1}^{n} y_i][\sum_{i=1}^{n} x_i]}{\sqrt{[n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2][n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2]}} \quad (6)$$

where, y_i is the value for the *ith* predicted attribute, x_i is the value for the *ith* measured attribute, and n represents the number of attributes.

The values for the correlation coefficients range from −1 (a perfect decreasing linear relationship) to +1 (a perfect increasing linear relationship). The absolute value for the coefficient indicates the strength of the relationship, with larger absolute values indicating stronger relationships [30].

In addition to the correlation coefficient, Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and Coefficient of Residual Mass (CMR) were used to evaluate the model.

$$MAE = \frac{\sum_{i=1}^{n} |x_i - y_i|}{n} \quad (7)$$

$$RMSE = [\frac{\sum_{i=1}^{n} (x_i - y_i)^2}{n}]^{0.5} \quad (8)$$

$$CRM = \frac{(\sum_{i=1}^{n} x_i) - (\sum_{i=1}^{n} y_i)}{\sum_{i=1}^{n} x_i} \quad (9)$$

The RMSE value indicates how much the model under- or over-estimates the measurements, and the CRM value is a measure of the tendency of the model to overestimate or underestimate the measurements. Positive values for CRM indicate that the model underestimates the measurements, and negative values for CRM indicate a tendency to overestimate. For a perfect fit between the observed and predicted data, the values for MAE, RMSE, and CRM should equal 0.0 [31].

5. Results and discussion

According to the calculated SPI values for different time scales, the k-nearest neighbor model was utilized to predict the most likely drought occurrence for the studied region during different years. In the beginning of computations, the optimum value for the K was obtained by the two-fold cross-validation method. Fig. 6 shows the precision of the method based on the Sum of Squares Error (SSE) coefficient. According to Fig. 6, three K values (15, 16, and 19) produced the same lowest error. The SSE value equal to 19 was selected because larger K values often smooth the K-NN model, thereby minimizing the risk of over-fitting. Then the most likely drought status for the region was predicted by the K-NN model. Fig. 7 shows the region drought status based on the standard precipitation index during the desired time period.

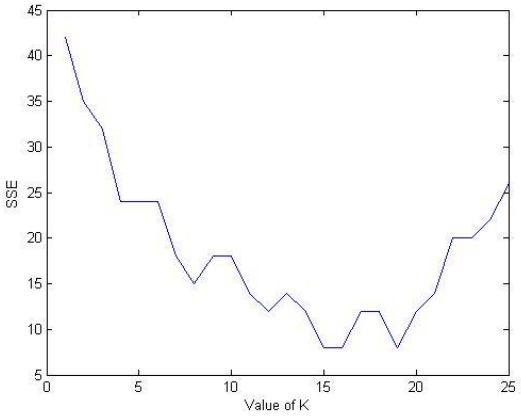

Figure 6. Two-fold cross-validation error rate for K-NN model.

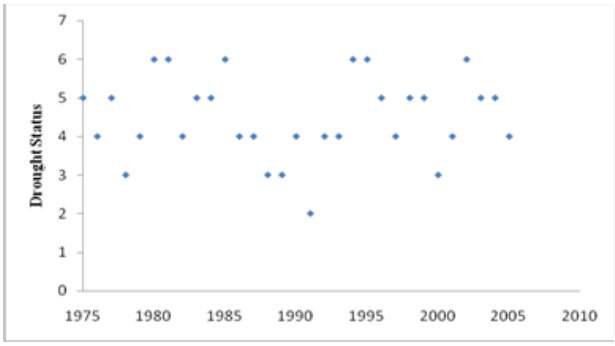

Figure 7. Region drought status based on SPI.

According to the results of the *K*-NN model, it can be found that the studied region has faced droughts over the years. Moreover, according to the presented moving time series, the standard precipitation index can be estimated for future years, and the most likely drought status can be determined.

In order to quantify the prediction accuracy and precision of the model, the Pearson correlation coefficient (r), mean absolute error (MAE), root mean square error (RMSE), and coefficient of residual mass (CRM) were calculated (Table 2). A high value for the Pearson correlation coefficient indicates strong relationships between the variables, and the low MAE, RMSE and CRM values show a reasonable precision and a low error of the *k*-NN model.

Comparing the results obtained for the K-NN modeling with the other SPI-based studies including Cancelliere et al. [11] (r = 0.715, MAD = 0.551 and RMSE = 0.731) indicates that the presented model gives appropriate predictions of the drought situation. Moreover, different time scales were considered in the model so that the drought predictions can be more reliable and efficient.

Table 2. Evaluation of K-NN model by some statistical coefficients.

r	MAE	RMSE	CRM
0.874	0.106	0.119	0.0011

6. Conclusion

Given the importance of drought monitoring in managing this phenomenon as well as the design and management of natural resources, water resource system planning, and various sectors of agriculture, in this study, using the standard precipitation index and *K*-NN model, a method was developed to predict drought occurrence. The model was evaluated using the precipitation and meteorological data of the city of Kerman, Iran. The results obtained indicate that this region has faced moderate-to-severe droughts for many years, which is consistent with the local observations. Finally, the efficiency and accuracy of the proposed model was evaluated by some statistical coefficients. The reasonable values for the Pearson correlation coefficient (r = 0.874), mean absolute error (MAE = 0.106), root mean square error (RMSE = 0.119), and coefficient of residual mass (CRM = 0.0011) indicate that the developed model is suitable and efficient.

References

[1] Mishra, A. K. & Singh, V. P. (2010). A review of drought concepts. Journal of Hydrology, vol. 391, pp. 202–216.

[2] Moreira, E. E., Coelho, C. A. & Paulo, A. A. (2008). A SPI-based drought category prediction using log linear models. Journal of Hydrology, vol. 354, pp. 116–130.

[3] Palmer, W. C. (1968). Keeping track of crop moisture conditions, nationwide: the new crop moisture index. Weatherwise, vol. 21, pp. 156–161.

[4] Van Rooy, M. P. (1965). A rainfall anomaly index independent of time and space. Notos, 14, 43.

[5] Gibbs, W. J. & Maher, J. V. (1967). Rainfall Deciles as Drought Indicators. Bureau of Meteorology Bull. 48. Commonwealth of Australia, Melbourne, Australia.

[6] Bhalme, H. N. & Mooley, D. A. (1980). Large-scale droughts/floods and monsoon circulation. Mon. Weather Rev, vol. 108, pp. 1197–1211.

[7] Shafer, B. A. & Dezman, L. E. (1982). Development of a Surface Water Supply Index (SWSI) to Assess the Severity of Drought Conditions in Snowpack Runoff Areas. In: Preprints, Western SnowConf., Reno, NV, Colorado State University, pp. 164–175.

[8] Gommes, R. & Petrassi, F. (1994). Rainfall Variability and Drought in Sub-Saharan Africa Since 1960. Agro-meteorology Series Working Paper 9, Food and Agriculture Organization, Rome, Italy.

[9] McKee, T. B., Doesken, N. J. & Kleis, J. 1993. The Relationship of Drought Frequancy and Duration to Time Scales, Eighth Conference on Applied Climatology. 17-22 January, Anaheim, California.

[10] Weghorst, K. M. (1996). The Reclamation Drought Index: Guidelines and Practical Applications. Bureau of Reclamation, Denver, CO, p. 6 (Available from Bureau of Reclamation, D-8530, Box 25007, Lakewood, CO 80226).

[11] Cancelliere, A., Di Mauro, G., Bonaccorso, B. & Rossi, G. (2007). Drought forecasting using the standardized precipitation index. Water resources management, vol. 21, no. 5, pp. 801-819.

[12] Han, P., Wang, P. X., Zhang, S. Y. & Zhu, D. H. (2010). Drought forecasting based on the remote

sensing data using ARIMA models. Mathematical and Computer Modelling, vol. 51, no. 11, pp.1398-1403.

[13] Farokhnia, A., Morid, S. & Byun, H. R. (2011). Application of global SST and SLP data for drought forecasting on Tehran plain using data mining and ANFIS techniques. Theoretical and applied climatology, vol. 104, no. (1-2), pp.71-81.

[14] Du, L., Tian, Q., Yu, T., Meng, Q., Jancso, T., Udvardy, P. & Huang, Y. (2013). A comprehensive drought monitoring method integrating MODIS and TRMM data. International Journal of Applied Earth Observation and Geoinformation, vol. 23, pp.245-253.

[15] Farahmand, A. & AghaKouchak, A. (2015). A generalized framework for deriving nonparametric standardized drought indicators. Advances in Water Resources, vol. 76, pp.140-145.

[16] Hao, C., Zhang, J. & Yao, F. (2015). Combination of multi-sensor remote sensing data for drought monitoring over Southwest China. International Journal of Applied Earth Observation and Geoinformation, vol. 35, pp.270-283.

[17] Narasimhan, B. & Srinivasan, R. (2005). Development and evaluation of Soil Moisture Deficit Index (SMDI) and Evapotranspiration Deficit Index (ETDI) for agricultural drought monitoring. Agricultural and Forest Meteorology, vol. 113, pp. 69–88.

[18] Türkeş, M. & Tatlı, H. (2009). Use of the standardized precipitation index (SPI) and a modified SPI for shaping the drought probabilities over Turkey. International Journal of Climatology, vol. 29, no. 15, pp. 2270-2282.

[19] Vasiliades, L., Loukas, A. & Liberis, N. (2011). A water balance derived drought index for Pinios River Basin, Greece. Water Resources Management, vol. 25, no. 4, pp.1087-1101.

[20] Dhaliwal, D. S., Sandhu, P. S. & Panda, S. N. (2011). Enhanced K-Nearest Neighbor Algorithm, World Academy of Science Engineering and Technology Journal, vol. 49, pp. 681-685.

[21] Mucherino, A., Papajorgji, P. & Pardalos, P. M. (2009). Data Mining in Agriculture, Springer.

[22] Smith, B. L., Williams, B. M., & Oswald, R. K. (2002). Comparison of parametric and nonparametric models for traffic flow forecasting, Transportation Research, vol. 10, no. 4, pp. 303–321.

[23] Prairie, J. R., Rajagopalan, B., Fulp, T. J., & Zagona, E. A. (2006). Modified K-NN model for stochastic streamflow simulation, Journal of Hydrologic Engineering, vol. 11, no. 4, pp. 371–378.

[24] Salas, J. D. & Lee, T. (2010). Nonparametric Simulation of Single-Site Seasonal Streamflows, Journal of Hydrologic Engineering, vol. 15, no. 4, pp. 284–296.

[25] Travis, Q. B., & Mays, L. W. (2011). Prediction of Intake Vortex Risk by Nearest Neighbors Modeling, Journal of Hydraulic Engineering, vol. 126, no. 5, pp. 701–705.

[26] Fadaei-Kermani., E, Barani, G. A., & Ghaeini-Hessaroeyeh., M. (2015). Prediction of cavitation damage on spillway using K-nearest neighbor modeling. Water science and technology. vol. 71, no. 3, pp. 347–352.

[27] Hastie, T., Tibshirani, R., & Friedman, J. (2008). The Elements of Statistical Learning, Second edition, Springer series, California.

[28] Xindung, W. & Kumar, V. (2009). Top Ten Algorithm in Data Mining, First edition, Taylor & Francis Group, USA.

[29] Bokharaeian, B. & Diaz, A. (2016). Extraction of Drug-Drug Interaction from Literature through Detecting Linguistic-based Negation and Clause Dependency. Journal of AI and Data Mining, vol. 4, no. 2, pp. 203-212.

[30] Izakian, Z. & Mesgari, M. (2015). Fuzzy clustering of time series data: A particle swarm optimization approach. Journal of AI and Data Mining, vol. 3, no. 1, pp. 39-46.

[31] Dashtaki., S. G, Homaee, M., & Mahdian., M. H. (2009). Site-Dependence Performance of Infiltration Models, Water Resour Manage, vol. 23, pp. 2777–2790.