Düzce University
Journal of Science & Technology

Research Article

Drought Estimation of Çanakkale with Data Mining

Özlem TERZİ [a,*], E. Dilek TAYLAN [b], Onur ÖZCANOĞLU [a], Tahsin BAYKAL [c]

[a] *Civil Engineering Department, Faculty of Technology, Süleyman Demirel University, Isparta, TÜRKİYE*
[b] *Civil Engineering Department, Faculty of Engineering, Süleyman Demirel University, Isparta, TÜRKİYE*
[c] *Water Management Department, Graduate School of Natural and Applied Sciences, Süleyman Demirel University, Isparta, TÜRKİYE*
** Corresponding author's e-mail address: ozlemterzi@sdu.edu.tr*

ABSTRACT

Drought estimation is important considering the harmful effects of the climate change in recent years. In this study, various models are developed with data mining technique for the drought estimation of Çanakkale, Turkey. Standardized precipitation index (SPI) values for 3, 6, 9, 12 and 24 months are calculated using the precipitation data of Çanakkale, Gökçeada and Bozcaada stations. The calculated SPI values of Gökçeada and Bozcaada are used as input parameters in developing data mining models with different algorithms. Examining the model results, it is observed that data mining technique is effective in drought estimation.

Keywords: *Drought, Data mining, SPI, Çanakkale, Turkey*

Veri Madenciliği ile Çanakkale İli Kuraklık Tahmini

ÖZET

Son yıllardaki iklim değişikliğinin zararlı etkileri göz önüne alındığında kuraklık tahmini oldukça önemlidir. Bu çalışmada, Çanakkale iline ait kuraklık tahmini için veri madenciliği ile modeller geliştirilmiştir. Çanakkale, Gökçeada ve Bozcaada yağış istasyonlarına ait yağış verileri ile 3, 6, 9, 12 ve 24 aylık standart yağış indeksi (SYİ) değerleri hesaplanmıştır. Hesaplanan Gökçeada ve Bozcaada'nın SYİ değerleri veri madenciliği modellerinde girdi olarak kullanılmıştır ve farklı algoritmalar ile modeller geliştirilmiştir. Model sonuçları, hesaplanan SYİ değerleri ile karşılaştırıldığında, veri madenciliği yönteminin kuraklık tahmininde iyi sonuçlar verdiği gözlemlemiştir.

Anahtar Kelimeler: *Kuraklık, Veri madenciliği, SYİ, Çanakkale, Türkiye*

Geliş: 31/05/2018, Düzeltme: 30/07/2018, Kabul: 06/09/2018

I. INTRODUCTION

Drought, which is a result of the increase in worlds population, urbanization, climate change, deforestation and desertification, has been threatening the environment and countries. Droughts have economic and social aspects. It is related with the economy, health and psychology of the society. Although negative effects of drought have become more and more obvious all over the world, the scope of drought is not yet fully understood and the effects are not adequately assessed [1]. Because of all these reasons, lots of studies have been conducted related to drought estimation. Dahal et al. [2] used SPI method at different time scales to study the temporal and spatial distribution of drought in central Nepal. Monthly rainfall data of 40 meteorological observation stations between the years 1981 and 2012 were used in the study. According to the results of the study, although there is no significant difference in regional rainfall trends, it is seen that there is a large interannual variation. The drought index trend analysis shows that the intensity and frequency of drought increases for most stations and that this trend is more powerful for longer periods of drought. During the study period, the most severe droughts were observed in the summer of 2004, 2005, 2006, 2009 and in the winter of 2006, 2008 and 2009. It is emphasized that these arid periods have a serious influence on the agriculture and animal husbandry in central Nepal. Gocic and Trakovic [3] evaluated the spatial and temporal characteristics of the drought in Serbia using monthly rainfall data from 29 stations between the years of 1948 and 2012. They applied the Percent of Normal Index (PNI) method to show the driest years in Serbia. In order to capture the drought patterns, they used the principal component analysis (PCA) and Standardized Precipitation Index (SPI). They applied the Agglomerative Hierarchical Cluster Analysis to determine 3 different drought sub-regions. Drought characteristics were analyzed at both country level and three sub-regions in terms of temporal variation of 12-month SPI values and drought frequency. It is observed that the monthly precipitation amounts are below the average of Serbia in R1 and R3 sub-regions, and above the average in the R2 sub-region. Within the observed period, the year 2000 is reported as the driest year and the year 1955 as the wettest year. With the advances in technology, artificial intelligence techniques are frequently used in the field of hydrology. It is seen that there are numerous studies using artificial intelligence techniques in literature [4-9].

Terzi [10] used data mining technique to develop different models for the rainfall estimation of Isparta, Turkey. Rainfall data of Senirkent, Uluborlu, Eğirdir and Yalvaç were used as input in the modelling stage. The most suitable model was obtained by the multilinear regression algorithm. Jalalkamali et al. [11] used support vector machine (SVM), multilayer perceptron artificial neural network (MLP-ANN), the autoregressive integrated moving average (ARIMAX) and adaptive neuro-fuzzy inference systems (ANFIS) model to estimate drought. Standardized precipitation index (SPI) values for 3 and 6 months (short-term) and 9, 12, 18 and 24 months (long-term) periods were estimated using the rainfall data of Yazd Rainfall Station for 51 years. SPI values of the years 1961-2002 were selected as training set and the remaining data of the years 2003-2012 were used in test set. As a result, it is seen that the ARIMAX model gave more appropriate results than the other models. Tadesse et al. [12] used data mining algorithms in order to identify the relationships between drought indices and oceanic indices. Based on the SPI and Palmer Drought Severity Index (PDSI), they determined the drought categories. They suggested that it can be monitored drought using oceanic indices as an indicator of drought by using data mining techniques. Choubin et al. [13] presented a drought index modeling approach by using the ANFIS, the M5P model tree and the multilayer perceptron (MLP) for the Maharlu-Bakhtegan basin located in southwestern Iran. They predicted the SPI values and demonstrated that performance of the MLP was better than the other models. Deo and

Şahin [14] developed models for drought estimation by using ANN algorithm for eight stations in eastern Australia. They predicted standardized precipitation and evapotranspiration index (SPEI). They said that ANN model was a beneficial tool for estimating monthly SPEI. Wu et al. [15] proposed a crisp distributed support vectors regression (CDSVR) model for monthly streamflow estimation. The performance of CDSVR model was compared with autoregressive moving average (ARMA), K-nearest neighbors (KNN), ANN, and crisp distributed ANN (CDANN) models. They stated that models developed by preprocessed data performed better than those developed by original data, and CDSVR model has the best performance compared to other models. Terzi and Baykal [16] used data mining process to estimate the amount of suspended sediment in the Kızılırmak river. They developed various sediment models by using river flow values as input parameters. The best model was obtained by the algorithm of M5'Rules, with the determination coefficient as 0.66. The results showed that data mining process could be used to estimate the amount of suspended sediment in river.

In this study, the validity of data mining technique in drought estimation is investigated. For this purpose, drought series are calculated by Standard Precipitation Index (SPI) for Gökçeada and Bozcaada stations located in the province of Çanakkale, Turkey. Subsequently, the calculated drought series are modeled using data mining technique.

II. Materials And Methods

A. STANDARDIZED PRECIPITATION INDEX (SPI)

Standardized Precipitation Index (SPI) is determined by the ratio between the standard deviation and the mean obtained from the long-term observed averages [17]. The SPI method is more flexible in examining the short and long periods of drought due to its application with long-term observations [18,19] SPI is developed to digitize precipitation for different time periods. For example, while soil moisture responds to precipitation differences on a relatively short time periods; groundwater, stream and reservoir storage features show rainfall conditions on longer time periods. For this reason, McKee et al. [17] calculated SPI for different time periods such as 3, 6, 12, 24 and 48 months.

SPI value is calculated by dividing the difference between the amount of precipitation (Xi) and the average of the amount of precipitation (Xmean) at a given time in a selected station by the standard deviation (σ) and given in Eq. 1 [20].

$$SPI=(x_i-x_{mean})/\sigma \quad (1)$$

While positive SPI values represent greater precipitation than average, negative values represent precipitation smaller than the average. McKee et al. [17] used a classification system to describe drought severity with SPI, as shown in the Table 1. They also define a criterion for drought event for any time period. The drought occurs when the SPI is consistently negative and reaches a density of -1.0 or less. The drought ends when the SPI reaches a value larger than “0” [21].

Table 1. SPI classification system [17]

SPI	Drought Category
2≤SPI	Extremely wet
1.50≤ $SPI \leq$ 1.99	Very wet
1.00≤ $SPI \leq$ 1.49	Moderately wet
-0.99≤ $SPI \leq$ 0.99	Near normal
-1.49≤ $SPI \leq$ -1.00	Moderately dry
-1.99≤ $SPI \leq$ -1.50	Severely dry
SPI≤-2	Extremely dry

B. DATA MINING

Data mining is a recent technology with great potential for identifying the most important information in databases. It is part of a larger process called knowledge discovery. Essentially, data mining discovers hidden relationships and patterns within large amounts of data. Data mining may be considered as advances in modeling techniques and statistical analysis to find relationships and useful patterns [22].

Data mining techniques can be grouped to two categories: (1) descriptive data mining, where the data are identified in accordance with their general features, (2) predictive data mining, where inference is performed from the current data for making predictions. Frequent patterns and association rules, clustering and deviation detection fall into the first group, whereas regression and classification fall into the other one [23].

Data mining (DM) process mainly consists of five stages, which are data understanding, data preparation, modeling, evaluation and knowledge as shown in Fig. 1.

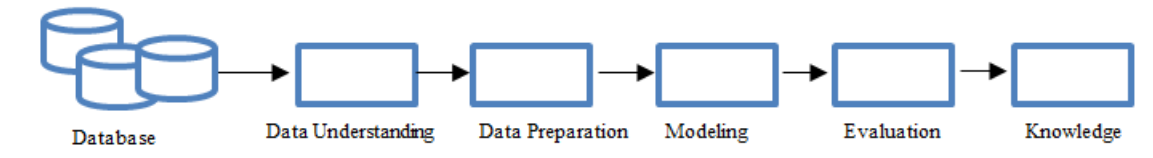

Figure 1. Data mining process

- Data understanding stage covers (1) data accumulation, (2) initial analysis of data and (3) detection of data quality problems.
- Data preparation stage includes (1) removing noise and inconsistent data, and (2) transformation for extracting the hidden features to make data understandable and meaningful.
- In the modeling phase, different modeling techniques are applied and the most suitable technique is found to achieve the objectives.
- The evaluation phase evaluates adequacy of the model obtained in the previous stage according to the selected performance criteria [24-26].

There are a lot of data mining algorithms. The explanation of algorithms only having appropriate results in drought modeling in this study is given below.
Random Subspace: This method creates a classifier, which is based on decision tree, in order to maintain the highest degree of accuracy in training data and to improve the accuracy in generalization as it grows in complexity. The classifier is made up of multiple trees created systematically by pseudo randomly selecting subsets of components of the feature vector, that is, trees constructed in randomly chosen subspaces [27].

KStar: K* is an instance-based classifier, that is the class of a test instance is based upon the class of those training instances similar to it, as determined by some similarity function. Using an entropy-based distance function, it differs from other instance-based learners [28].

Bagging: Bootstrap Aggregation or Bagging for short is an ensemble algorithm that can be used for classification or regression. Bootstrap is a statistical prediction method where a statistical quantity like a mean is predicted from multiple random samples of the data (with replacement). It is a beneficial method in case of a need for a more robust estimate of a statistical quantity with a limited amount of data [29].

LeastMedSq: This method applies a least median squared linear regression making use of the existing weka LinearRegression class to make predictions. Random subsamples of the data are used to create least squared regression functions. The least squared regression with the lowest median squared error is selected as the final model [30].

C. STUDY AREA AND DATA

Located between 25°35'-27°45'E and 39°30'-40°45'E in the northwest of Turkey, Çanakkale province covers a surface area of 9737 km^2 on both Europe and Asia [31]. Çanakkale has a semi-humid climate with an average total annual rainfall of 591.5mm for long years. The maximum amount of rainfall measured in 24 hours is 137.8 mm [32]. Besides the city center of Çanakkale, the study area covers also two islands on the Aegean Sea, which are Gökçeada and Bozcaada, as shown in Fig. 2.

In this study, the monthly total precipitation data of Çanakkale, Gökçeada and Bozcaada stations for the years 1975-2010 taken from the Turkish State Meteorological Service have been used. Before the modelling stage, %80 of the whole data has been allocated to training set. The remaining part of the data has been used in testing set.

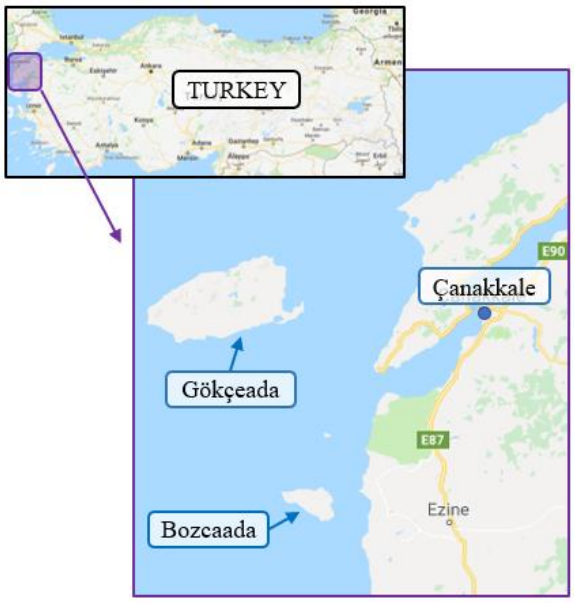

Figure 2. Study area

III. Results And Discussion

Prior to the development of the drought models, the homogeneity of the data was tested by the double mass curve method and it was seen to be homogeneous of three stations. Then, the SPI values for 3, 6, 9, 12 and 24 months have been calculated using the rainfall data of Çanakkale, Gökçeada and Bozcaada stations. The calculated SPI values of Bozcaada and Gökçeada stations are used as inputs in data mining (DM) models and drought series of Çanakkale station are estimated.

Models have been developed using different types of algorithms in DM techniques. It was determined performance of the models using determination coefficients (R^2), mean absolute error (MAE) and root mean squared error (RMSE) in Eqs. (6), (7) and (8).

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(X_{i(observed)} - X_{i(model)}\right)^2}{\sum_{i=1}^{N}\left(X_{i(observed)} - X_{i(mean)}\right)^2} \quad (6)$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|X_{i(observed)} - X_{i(model)}\right| \quad (7)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(X_{i(observed)} - X_{i(model)}\right)^2} \quad (8)$$

Where N is the number of observed data, $X_{i(observed)}$ and $X_{i(model)}$ are SPI values and DM results, respectively.

To develop models, WEKA software is used in this study. All the methods in this program which are usable and suitable for modeling have been tested. Only the models having best results for each time period (3-, 6-, 9-, 12-, 24-months) are mentioned here. The best results have been obtained with RandomSubSpace, LeastMedSqand Bagging algorithms for 3-, 6- and 12-month models, respectively.

Also, Kstar algorithm is found to be appropriate for 9- and 24-month models. The R^2, RMSE and MAE values for the best model results are given in Table 2.

Table 2. *Performance of the models*

Drought Period	Algorithms	Training Set			Test Set		
		RMSE	MAE	R^2	RMSE	MAE	R^2
3-month	RandomSubSpace	0.537	0.002	0.589	0.569	0.014	0.514
6-month	LeastMedSq	0.480	0.043	0.737	0.381	0.012	0.808
9-month	Kstar	0.407	0.012	0.833	0.406	0.070	0.770
12-month	Bagging	0.403	0.012	0.839	0.417	0.050	0.805
24-month	Kstar	0.436	0.067	0.827	0.420	0.033	0.785

When the results are examined, quite remarkable performance is observed for estimating the 6, 9, 12 and 24-month SPI values. On the other hand, it is seen that models are not very successful in estimating the 3-month SPI values. Also, it is seen from scatter diagram of 3-month models given in Fig. 3 and time series given in Fig. 4.

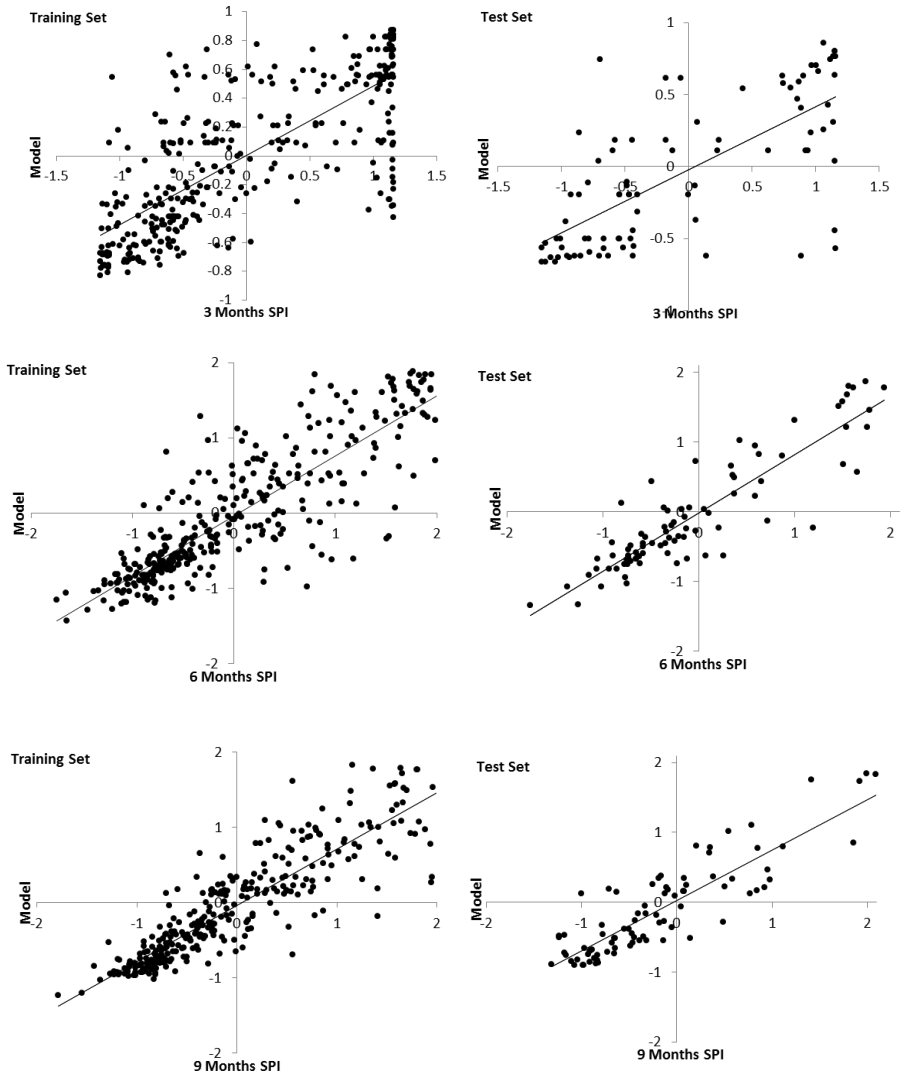

Figure 3. *Scatter diagrams of the best models*

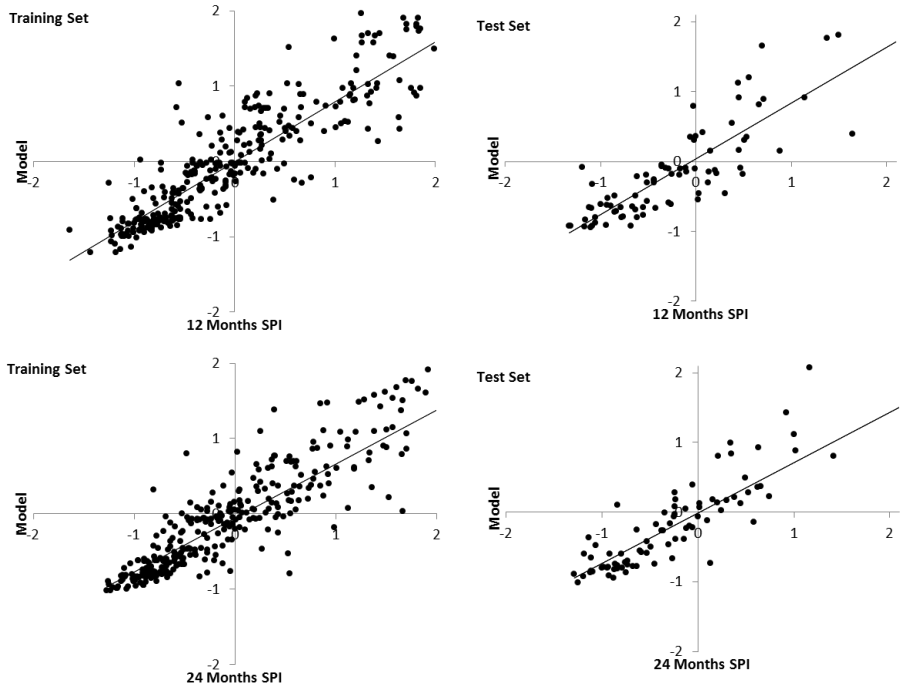

Figure 3. (continue). *Scatter diagrams of the best models*

Fig. 3 shows that the 3- month model gives inadequate and untidy results. Also, it is shown that 3-month model could not estimate extreme values in Fig. 4. The other models are generally around fit lines for training and testing sets. According to all evaluation criteria, scatter diagrams and time series, it is found that 6-month model has the highest performance.

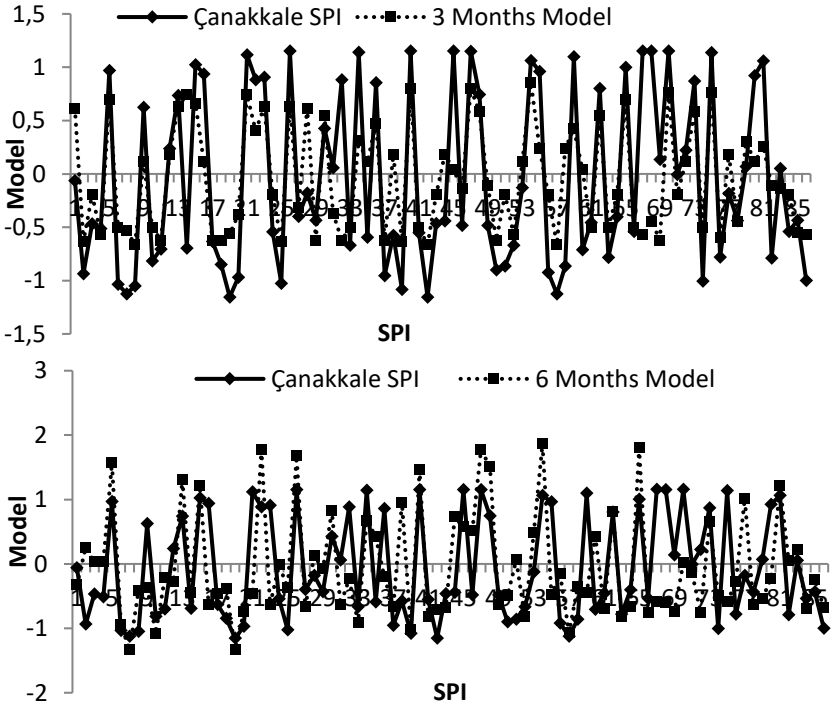

Figure 4. Time series of the models

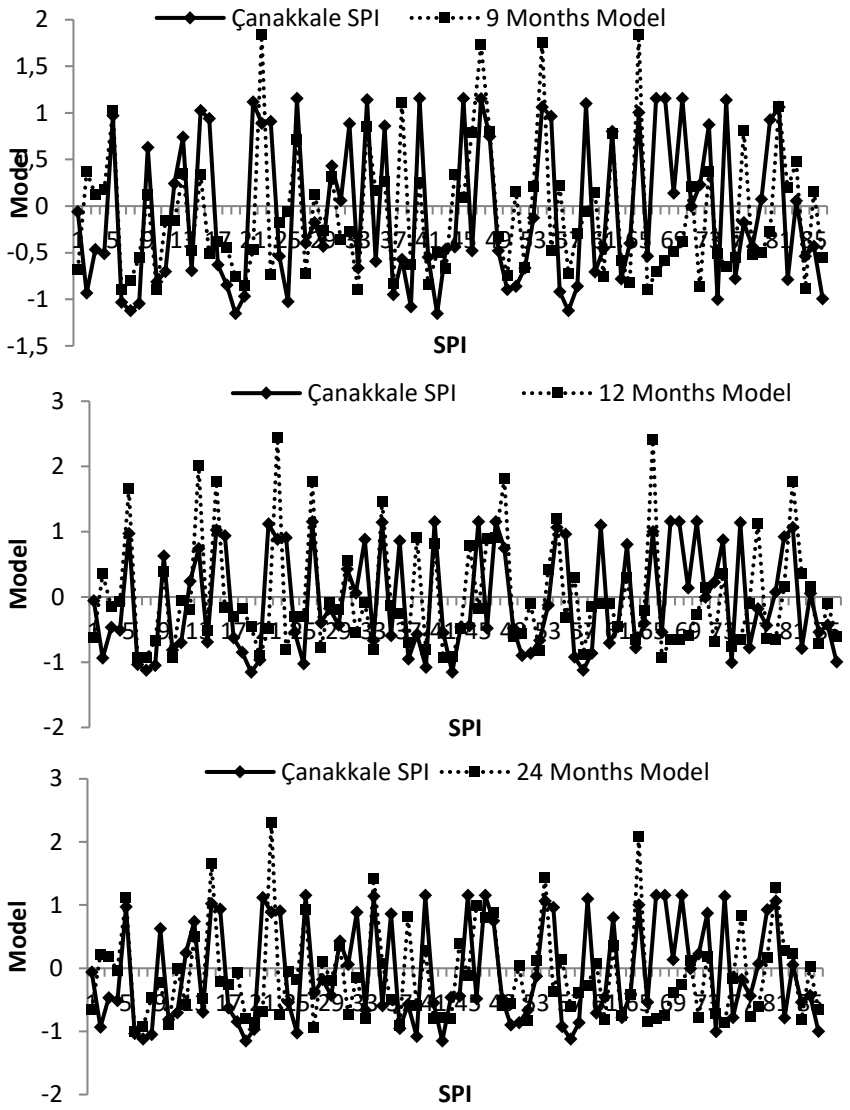

Figure 4. *(continue).* Time series of the models

IV. CONCLUSIONS

In this study, various drought models have been developed for Çanakkale Province using data mining technique. Firstly, SPI values of Gökçeada, Bozcaada and Çanakkale stations for 3, 6, 9, 12 and 24 months have been calculated. These calculated SPI values have been used as input parameters while developing data mining models. The DM models were developed using different algorithms. When the results of the developed models are analyzed, it is seen that 3-month SPI model shows relatively poor performance. However, the results of the models for 6, 9, 12 and 24 months are quite remarkable. It can be seen that 6-month model results and SPI values have a higher agreement than the other models, especially. In conclusion, data mining technique seems to be useful in drought estimation.

V. REFERENCES

[1] S. Sırdaş and Z. Şen, "Meteorological drought modelling and application to Turkey," *Itu Journal/D Engineering*, vol. 2, no.2, pp. 95-103, 2003.

[2] P. Dahal, N. S. Shrestha, M. L. Shrestha, N. Y. Krakauer, J. Panthi, S. M. Pradhanang, A. Jha and T. Lakhankar, "Drought risk assessment in central Nepal: temporal and spatial analysis," *Natural Hazards*, vol. 80, no. 3, pp. 1913-1932, 2016.

[3] M. Gocic and S. Trajkovic, "Spatiotemporal characteristics of drought in Serbia," *Journal of Hydrology*, vol. 510, pp.110-123, 2014.

[4] H.Vathsala and S.G. Koolagudi, "Prediction model for peninsular Indian summer monsoon rainfall using data mining and statistical approaches," *Computers and Geosciences*, vol. 98, pp. 55-63, 2017.

[5] Ö. Terzi, E. D. Taylan, O. Özcanoğlu and T. Baykal, "A Wavelet-ANFIS Hybrid model for drought forecasting," 8th International Advanced Technologies Symposium, Elazığ, Turkey, 2017, pp. 3251-3258.

[6] S. P. Norman, F. H. Koch and W. W. Hargrove, "Review of broad-scale drought monitoring of forests: Toward an integrated data mining approach," *Forest Ecology and Management*, vol. 380, pp. 346-358, 2016.

[7] H. A. Afan, A. El-shafie, W. H. M. W Mohtar and Z. M.Yaseen, "Past, present and prospect of an Artificial Intelligence (AI) based model for sediment transport prediction," *Journal of Hydrology*, vol. 541, pp. 902-913, 2016.

[8] M. Zounemat-Kermani, Ö. Kişi, J. Adamowski and A. Ramezani-Charmahineh, "Evaluation of data driven models for river suspended sediment concentration modeling," *Journal of Hydrology*, vol. 535, pp. 457-472, 2016.

[9] V. Nourani, A.H. Baghanam, J. Adamowski and Ö. Kisi, "Applications of hybrid wavelet–artificial intelligence models in hydrology: A review," *Journal of Hydrology*, vol. 514, pp. 358-377, 2014.

[10] Ö. Terzi, "Veri madenciliği süreci kullanilarak yağiş tahmini," Akıllı Sistemlerde Yenilikler ve Uygulamaları Sempozyumu, Trabzon, Turkey, 2012, ss. 126-129.

[11] A. Jalalkamali, M. Moradi and N. Moradi, "Application of several artificial intelligence models and ARIMAX model for forecasting drought using the Standardized Precipitation Index," *International Journal of Environmental Science and Technology*, vol. 12, no. 4, pp. 1201-1210, 2015.

[12] T. Tadesse, D. A. Wilhite, S. K. Harms, M. J. Hayes and S. Goddard, "Drought monitoring using data mining techniques: A case study for Nebraska, USA," *Natural Hazards*, vol. 33, no. 1, pp. 137-159, 2004.

[13] B. Choubin, A. Malekian, and M. Golshan, "Application of several data-driven techniques to predict a standardized precipitation index." *Atmósfera*, vol. 29, no. 2, pp. 121-128, 2016.

[14] R. C. Deo and M. Şahin, "Application of the artificial neural network model for prediction of monthly standardized precipitation and evapotranspiration index using hydrometeorological parameters and climate indices in eastern Australia," *Atmospheric Research*, vol. 161, pp. 65-81, 2015.

[15] C. L. Wu, K. W. Chau and Y. S. Li, "Predicting monthly streamflow using data-driven models coupled with data-preprocessing techniques," *Water Resources Research*, vol. 45, no. 8, 2009.

[16] Ö.Terzi and T. Baykal, "Data mining process for river suspended sediment estimation," *SDU International Journal of Technological Science*, vol. 8, no 3, pp. 19-26, 2016.

[17] T. B McKee, N. J. Doesken and J. Kleist, "The relationship of drought frequency and duration to time scales," Eighth Conference on Applied Climatology, American Meteorological Society, Anaheim, CA, 1993, pp. 1-6.

[18] D. C. Edwards and T. B. McKee, "Characteristics of 20th century drought in the united states at multiple time scales," *Atmospheric Science Paper*, vol. 634, pp. 1-30, 1997.

[19] K. T. Redmond, "Integrated climate monitoring for drought detection," Drought: A Global Assessment, edited by Wilhite, DA, Routledge, London, 2000.

[20] D. Atmaca , "Regional drought analysis on Konya province by using standardized precipitation index (SPI)," M.S. Thesis, Department of Agricultural Structures and Irrigation, Gaziosmanpaşa University, Tokat, Turkey, 2011.

[21] WMO, "Standardized Precipitation Index User Guide," WMO-No. 1090, World Meteorological Organization, 2012.

[22] H. Edelstein, (1997). [Online] Available: http://wwww.db2mag.com/db_area/archives/19

[23] J. Han and M. Kamber, *Data mining: concepts and techniques*, 3rd ed., New York, USA: Elsevier, 2006, pp. 770.

[24] S. T. Li and L. Y. Shue, "Data mining to aid policy making in air pollution management," *Expert System and Applications*, vol. 27, pp. 331-340, 2004.

[25] Z. H. Zhou, "Three Perspectives of Data Mining", *Artificial Intelligence*, vol. 143, no.1, pp. 139–146, 2003.

[26] R. Mattison, Data Warehousing: Strategies, Technologies and Techniques Statistical Analysis, SPSS Inc. WhitePapers, 1996.

[27] T. K. Ho, "The random subspace method for constructing decision forests", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832-844, 1998.

[28] J. G. Cleary, L.E. Trigg, “K* an instance-based learner using an entropic distance measure”, In: 12th International Conference on Machine Learning, 1995, pp. 108-114.

[29] B. Jason, (2016, July 27). [Online]. Available: https://machinelearningmastery.com/use-ensemble-machine-learning-algorithms-weka/

[30] P. J. Rousseeuw and M.L. Annick, *Robust regression and outlier detection*, 1st ed., vol. 589. New Jersey, USA: John Wiley & Sons, 2005.

[31] Çanakkale Municipality. (2018, Feb 10). [Online]. Available: http://www.canakkale.bel.tr/icerik/1941/cografi-yapi

[32] İzmir Meteorology Directorate. (2018, Feb 10). [Online]. Available: http://izmir.mgm.gov.tr/FILES/iklim/canakkale_iklim.pdf