Atmósfera 25(1), 85-105 (2012)

Seasonal and annual regional drought prediction by using data-mining approach

K. YUREKLI
Department of Biosystem Engineering, Faculty of Agriculture, University of Gaziosmanpasa, 60240 Tasliciftlik, Tokat, Turkey

M. TAGHI SATTARI
Department of Water Engineering, Faculty of Agriculture, University of Tabriz, Tabriz 5166614766, Iran

A. S. ANLI
Department of Farm Structure and Irrigation, Agriculture Faculty, University of Ankara, Ankara-Turkey

M. A. HINIS
Department of Civil Engineering, Faculty of Engineering, University of Aksaray, Aksaray-Turkey
Corresponding author; e-mail: mhinis@gmail.com

Received March 3, 2010; accepted July 29, 2011

RESUMEN

Este estudio examina el análisis de la sequía estacional regional con base en el método del índice estandarizado de precipitación (SPI, por sus siglas en inglés) y en la técnica del árbol de decisiones que es una aproximación de minería de datos. Se formaron series de precipitación acumulada para cinco periodos de referencia (cuatro series estacionales y una anual) utilizando la precipitación mensual de 17 estaciones de la cuenca de Cekerek en Turquía, que tiene un área de 1 165 440 ha. Se realizó un análisis regional agrupando las estaciones inicialmente como grupos homogéneos de acuerdo con el criterio de discordancia considerando las tasas de momento-l. No hubo estaciones discordantes de acuerdo con las medidas de discordancia de las características de los sitios, excepto para las del primer período de referencia. Las medidas de heterogeneidad muestran que los grupos seleccionados fueron homogéneos. Con base en el criterio de bondad de ajuste |ZDIST| las distribuciones regionales candidato con |ZDIST| mínimo para los periodos de referencia k fueron la Pareto generalizada (GPA), la de valores extremos generalizados (GEV), la logística generalizada (GLO) la Pearson tipo III (PE·), la GEV y la log normal de 3 parámetros (LN3), respectivamente. Las categorías de sequía para cada región se predijeron aplicando el árbol de decisiones obtenido de la fase de entrenamiento para los periodos k de referencia. Los resultados revelan que no hubo diferencia significativa entre las categorías de sequía calculadas con el algoritmo convencional de SPI y las de la aproximación por el árbol de decisiones. Más aún, la exactitud de la predicción para los periodos de referencia k fue mayor que 94 %, excepto para los períodos de referencia k3 (81.2 %) y k5 (86.4 %).

ABSTRACT

This study examines the seasonal regional drought analysis based on the standardized precipitation index (SPI) method and the decision tree technique which is a data-mining approach. The cumulative rainfall series for five reference periods (four seasonal and one annual series) were constituted by using monthly rainfalls from 17 stations in Cekerek Watershed, Turkey, which has an area of 1165 440 ha. Regional analysis was performed by forming the stations initially as homogeneous group(s) according to the discordancy criteria considering by l-moment ratios. There was no discordant station according to discordancy measure of site characteristics except for the first reference period. The heterogeneity measures showed that the selected groups were homogeneous. Based on the goodness of fit criteria $|Z^{DIST}|$ the candidate regional distributions having the minimum Z^{DIST} for k-reference periods were the Generalized Pareto (GPA), Generalized Extreme Values (GEV), Generalized Logistic (GLO), Pearson Type III (PE3), GEV and 3-parameter Log Normal (LN3), respectively. The drought categories for each region were predicted by applying the decision tree rules obtained from the training phase of the k-reference periods. The results revealed that there was no significant difference between drought categories calculated from the conventional SPI algorithm and decision tree approaches. Moreover, the accuracy of prediction for k-reference periods was greater than 94%, except for k3 (81.2) and k5 (86.4%) reference periods.

Keywords: L-moments, regionalization, standard precipitation index, decision tree.

1. Introduction

Drought is one of the most serious problems for human societies and ecosystems arising from climate fluctuations and variations. Although its impact does not come through sudden events, such as floods and storms, drought is one of the most damaging types of natural disasters influencing for longer periods. Initiation of drought is less noticeable and there are no rapid physical disruptions at the beginning. However, droughts may become disastrous in time and spread into wide areas by affecting many more social, economical and environmental aspects than other types of disasters do. Drought can last for long time and sustain the impact for longer durations. Human interferences often increase the impact of drought because of a high use of water that cannot be supported when the natural supply is limited. Although it is not easy to define droughts precisely, they can be simply considered as periods of insufficient precipitation and water supply relative to average conditions, however, operational definitions may often help to define the onset, severity and end of droughts. Le Houerou (1996) stated that droughts were experienced in almost all types of agricultural land in the world, but arid lands are most susceptible.

Drought is classified as agricultural, hydrological or meteorological. Agnew and Warren (1996) described agricultural drought as a spatial phenomenon that causes significant reductions in agricultural productivity, mainly due to an inadequate supply of soil moisture. Hydrological drought refers to deficiencies in surface and subsurface water supplies (Palmer, 1965). Meteorological drought is usually measured by how far the precipitation from normal has been over a certain period of time (Agnew, 1990).

Numerous indices were designed to quantify agricultural, hydrological and meteorological droughts. Drought indices derived from hydroclimatical data are supposed to provide a concise information about the drought condition of a region. These indices are aften used for making decisions on water resources management and water allocations for minimizing the impact of drought. Researchers have focused on standardized precipitation index (SPI) recently to examine the problems such as drought, flood and crop yields. SPI quantifies the precipitation deficit and may be applied in areas with different climates for various time scales (Edwards and McKee,

1997). SPI is based on the monthly precipitation data summed at different time scales and fitted to a statistical distribution.

Loukas and Vasiliades (2004) examined the temporal and spatial characteristics of meteorological drought to provide a framework for sustainable water resources management in the region of Thessaly, Greece by using the SPI as an indicator of both the drought severity and the characteristics of droughts. Yamoah (2000) investigated the effects of the SPI and fertilizer nitrogen (N) rate on yield and risk of maize-based cropping systems in northeast Nebraska. They expressed that the SPI would be used as an indicator to choice of crops, N levels, and management decisions to conserve water in rainfed cropping systems. Selier (2002) used the SPI as a tool for monitoring flood risk affecting the southern Cordoba province in Argentina. Giddings (2005) implied that the SPI were used with notable success in various applications as an indicator of drought severity or excessive wetness. Alatise and Ikumawoyi (2007) applied four techniques namely, the Stochastic Component Time Series (SCTS), the Rainfall Anomaly Index (RAI), the Cumulative Rainfall Information (CRI) and the Drought Severity Index (DSI) to a 73-year rainfall data for the evaluation of drought in Lokoja, Nigeria. The RAI was selected as the most appropriate technique because of its ability to supply more information on drought occurrences in the study area more than the other three techniques. Oladipo (1985) examined the performances of three drought indices, namely the RAI, Bhalme and Mooley drought index (BMDI) and the Palmer drought index (PDI) and stated that the three indices appeared to be effective in detecting drought periods. Wu *et al.* (2004) developed an agricultural drought risk-assessment model for corn and soybeans by using the standardized precipitation index and crop-specific drought index. The 26-time scales of the SPI were included for this reason, and the SPI values at four time scales (4, 10, 32 and 52 weeks) were selected for model development. Labedzki (2007) estimated meteorological drought frequency in the region of Bydgoszcz in the central part of Poland by taking into consideration the SPI values at 3-, 6-, 12-, 24- and 48-month timescales.

In this study, regional drought analysis based on the SPI method and data mining approach were carried out. Large historical datasets are required to identify the complex inter-relationship between different climatic parameters and to distinguish patterns that may be used to predict drought. In this sense, an automated and efficient way is desired to extract reasonable information from such large data archives. This problem can be overcome by using data mining approach, which is a relatively new method developed for extracting relevance information from large datasets. Tadesse (2004) reported that data mining approach was used for commercial applications, medical research, and telecommunications, but it was not for drought analysis. Therefore, they analyzed the usability of the technique to find associations between drought and several oceanic and climatic indices, and suggested that data mining technique could be used to monitor drought. Sharma (2006) used the SPI and Vegetation Condition Index (VCI) as input parameters for generating the rules related to data mining, and concluded that data mining technique by using association rule and independent component analysis was successfully applied, and it was possible to extract information about the temporal and spatial pattern of drought. Belda and Penades (2007) examined data mining by using OLAP-mining technique to determine the association rules between synoptic patterns and climatic index.

Main objective of the present study is to perform seasonal and annual regional drought analysis based on SPI by means of decision tree technique which is a data-mining approach. The study was arranged in four consecutive stages as follows. The first step was to constitute the cumulative

rainfall series for the k-reference periods by using monthly rainfalls from 17 stations in Cekerek Watershed. The second stage was to form sub-homogeneous regions for the regional frequency analysis and to choose the best fit regional distribution for the cumulative rainfall series obtained from the stations in the sub-homogeneous regions. The third stage was to transform the cumulative rainfall series in the sub-homogeneous region to normal (Gaussian) symmetrical distribution by using the candidate regional distribution to find the z-score (SPI) relationship. The SPI classification suggested by McKee *et al.* (1993) is given in Table I. In the last stage, the decision tree technique was applied to the cumulative rainfall series to delineate drought categories based on the SPI values.

Table I. Standardized Precipitation Index Classification.

SPI Values	Classifications	Abbreviation
2.00 and more	Extremely wet	EW
1.50 to 1.99	Very wet	VW
1.00 to 1.49	Moderately wet	MW
0.99 to 0.00	Normal	N
0.00 to –0.99	Near normal	NN
–1.00 to –1.49	Moderately drought	MD
–1.50 to –1.99	Severe drought	SD
–2.00 and less	Extremely drought	ED

2. Materials and methods

2.1 Cekerek watershed

The Çekerek Stream watershed lies in between 39º 30’ and 40º 45’ N and 35º 15’ and 36º 15’ E. This area covers approximately 1165 440 ha, which is about 1.5% of Turkey’s total area. The study area is located on the north Anatolia fault line that is one of the most effective faults in the world. Therefore, tectonic movement affects the characteristics of the watershed. The Çekerek Stream is formed by the confluence of small streams that originate from the Kizik, Dinar, Çali and Kavak hills, near the Çamlibel district. The Çekerek Stream is approximately 276 km in length. The stream drains into the Yesilirmak River near Kayabasi (Anonymous, 1970). In this study, four seasonal (SRS) and one annual rainfall (ARS) series were formed by using monthly total precipitation series obtained from 17 selected rain gauge stations in the Cekerek watershed, Turkey. The selected 17 stations, managed by the Turkish State Meteorological Service and General Directorate of State Hydraulic Works, were scattered over the Cekerek watershed to represent fully the precipitation regimes affecting the area. The approximate locations of the rain gauge stations are shown in Figure 1. There is a lack of data in monthly total rainfalls of some years for some of the rain gauge stations in the studied region. The year of interest was discarded for the k-reference period with lack of the data. The data records of 17 stations were given in Table II.

2.2 Analysis of data

It is assumed that a time series of monthly rainfall depths, $P_{i,j}$, is available where *i* denotes the year and *j* denotes the month. The seasonal rainfall depth series for the k-th reference period is obtained as:

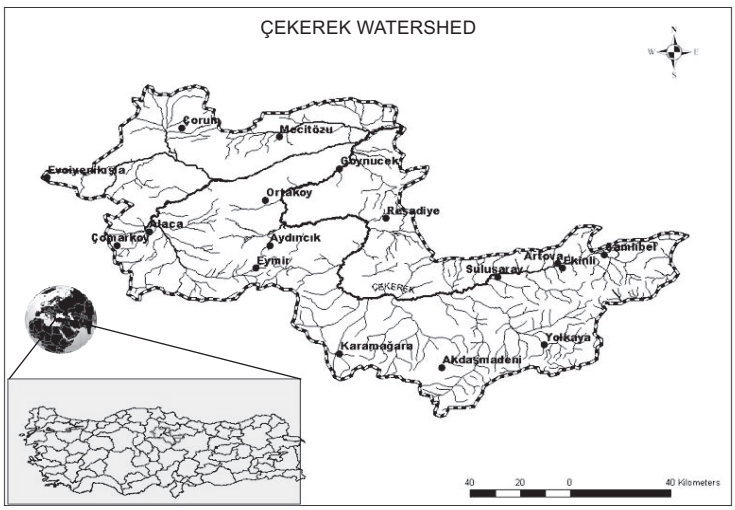

Fig. 1. Rainfall gauge station over Cekerek watershed.

Table II. Discordancy analysis results of rainfall gauge stations over the k1 (Jan-March) and k2 (April-June) reference periods.

	Stations	Record length	k1 reference period (January-March)				k2 reference period (April-June)			
			l-Cv	l-Cs	l-Ck	Di	l-Cv	l-Cs	l-Ck	Di
	Corum	1929-2006	0.1835	–0.0185	0.0428	1.25	0.1929	0.1056	0.111	0.44
	Eymir	1986-1996	0.1484	0.103	0.0738	1.14	0.1891	–0.0677	0.1636	0.33
	Ortakoy	1989-2007	0.1677	0.0203	0.0292	0.74	0.2027	0.0316	0.0442	0.86
k11	Alaca	1967-2007	0.2032	0.0144	0.0039	0.35	0.2	0.1088	0.1736	0.31
	Aydincik	1969-1990	0.1597	–0.0168	0.1506	2.12	0.2856	0.1572	0.1385	1.75
	ZResadiye	1968-2004	0.2137	0.0521	0.0044	0.81	0.2053	0.1587	0.1979	0.26
	Goynucek	1966-1988	0.2039	–0.0444	0.0018	0.57	0.1582	0.0865	0.2404	1.29
	Comarkoy	1966-1979	0.179	–0.0942	0.0117	1.03	0.1645	–0.0973	0.2964	2
	Mecitozu	1984-1998	0.1849	0.0732	0.2099	1.78	0.1921	–0.1224	0.1054	0.95
	Camlibel	1966-1976	0.2349	–0.0272	–0.0617	1.7	0.1934	–0.0406	0.0766	0.47
	Akdagmadeni	1964-1990	0.1702	0.2097	0.2003	1.24	0.1526	–0.1227	0.1075	0.6
k12	Ekinli	1967-1999	0.2211	0.1363	0.1499	0.21	0.2025	0.0224	0.0492	0.8
	Sulusaray	1966-2001	0.1733	0.0803	0.1219	0.09	0.1945	0.1785	0.1954	0.45
	Artova	1966-1990	0.2067	–0.0173	0.0804	1.65	0.1361	0.103	0.2316	1.63
	Karamagara	1959-1994	0.2833	0.0903	0.0066	0.88	0.295	0.3811	0.349	2.95
	Yolkaya	1979-1994	0.301	0.1254	0.0952	0.66	0.1872	–0.1188	0.1286	0.56
	Evciyenikisla	1970-2002	0.233	0.0757	0.0917	0.79	0.192	0.061	0.3153	1.37

l-Cv, the sample l-coefficient of variation
l-Cs, the sample l-coefficient of skewness
l-Ck, the sample l-coefficient of kurtosis
D_i, discordancy measure for site i
k11, first probable homogeneous region for k1 reference period
k12, second probable homogeneous region for k1 reference period

$$R_{i,k} = \sum_{j=3k-2}^{3k} P_{i,j} \qquad i = 1, 2, \ldots, N \qquad j = 1, 2, \ldots, 12 \qquad k = 1, 2, 3, 4 \tag{1a}$$

and annual rainfall depth series is obtained as

$$R_{i,k} = \sum_{j=1}^{12} P_{i,j} \qquad i = 1, 2, \ldots, N \qquad j = 1, 2, \ldots, 12 \qquad k = 5 \tag{1b}$$

where $R_{i,k}$ is the cumulative rainfall depth for the k-th reference period of *i*-th year, $k = 1$ for January-March, $k = 2$ for April-June, $k = 3$ for July-September, $k = 4$ for October-December and $k = 5$ for January-December (annual) time periods.

2.3 Standardized precipitation index (SPI) algorithm

The SPI developed by McKee (1993) is a way of measuring drought characteristics based only on precipitation data. The SPI is used to monitor conditions on a variety of time scales. Technically, the SPI is the number of standard deviations that the observed value would deviate from the longterm mean, for a normally distributed random variable. The SPI have some advantages for the following reasons. Precipitation is the only variable in the SPI calculation. Therefore, this index can be applied to any regions where the availability of climatic variables limits the use of other widely used indices such as Palmer Drought Index (PDI). To have a wide spectrum of time scales make SPI more flexible for both short-term and long-term drought monitoring than any other indices (Edwards and McKee, 1997; Redmond, 2000). Alley (1984) and Guttman (1998) compared SPI and PDSI, and spatial inconsistency was found in PDSI, thefore SPI was recommended for drought studies. SPI is comparable both in time and space and is not affected by geographical or topographical differences (Lana, 2001). The SPI algorithm is conceptually equivalent to z-scores commonly used in statistics:

$$\mathrm{SPI} = \frac{p_i - \sum_{i=1}^{n} p_i / n}{\sigma_p} \tag{2}$$

where, SPI represents the standardized precipitation index, p_i is the rainfall for a given period, *n* is the total length of record and σ_p is the standard deviation.

McKee (1993) used the drought classification system shown in Table I to define intensities resulting from the SPI. A drought event occurs any time when the SPI is continuously negative and reaches an intensity where the SPI is -1.0 or less. The event ends when the SPI becomes positive.

It is known that rainfall data is typically positively skewed. Therefore, the precipitation data should be transformed to a more normal or Gaussian symmetrical distribution to use the z-score relationship. McKee (1993, 1995) and Komuscu (1999) implied that the long-term rainfall data sets must be first normalized to determine the SPI of the data sets. The application of many researchers related to the transformation of monthly rainfall is the gamma distribution. Thom (1966) stated that monthly rainfall generally fit to the gamma distribution. Guttman (1999) examined impact

of six distributions on SPI and recommended that Pearson Type III distribution is the best way to normalize long-term data when calculating SPI. Edwards and McKee (1997) suggested gamma distribution with two parameters to transform the precipitation data. Kumar (2009) investigated the use of SPI for drought intensity assessments and found that SPI values calculated by gamma distribution underestimate dryness and wetness caused by very low and very high rainfall. Therefore they stated that there is a need to use other statistical distributions for SPI computation for improving the sensitivity.

Before executing the transformation, it is an important task to find the best distribution representing the precipitation data since it has an impact on the SPI. Therefore, it was decided to use the l-moment approach introduced by Hosking (1990) to choose the best fit regional distribution in the study.

2.3.1 l-Moment approach

The l-moments are first defined by Hosking (1990) as an alternative approach of describing the shape of probability distributions. They are analogous to conventional moments with measures of location, scale and shape, and able to be computed from linear combinations of order statistics. The l-moments have some theoretical advantages over conventional moments. These advantages are that they are mostly robust and less sensitive to outliers, so that l-moments are calculated as linear combination of the ordered data sequence unless squaring or cubing the data. Moreover, the parameter estimations are more reliable than the conventional method of moment estimates, particularly from small samples, and are usually computationally more tractable than maximum likelihood estimates. On the other hand, estimators of l-moments are virtually unbiased (Hosking and Wallis, 1997). Basically, l-moments are linear functions of probability weighted moments (PWMs). The PWMs are defined by Greenwood (1979) as;

$$ß_{rj} = E\left\{x_j\left[F_j(x_j)\right]^r\right\} \qquad j=1, 2,\ldots, n \tag{3}$$

Where $ß_{rj}$ is the r^{th} order PWM at site j and $F_j(x_j)$ is the cumulative distribution function (cdf) of x_j at site j. For any given site, the four first l–moments based on the PWMs are defined;

$$\lambda_1 = \beta_0, \qquad \lambda_2 = 2\,\beta_2 - \beta_1,$$
$$\lambda_3 = 6\,\beta_3 - 6\,\beta_2 + \beta_1, \qquad \lambda_4 = 20\,\beta_4 - 30\,\beta_3 + 12\,\beta_2 - \beta_1 \tag{4}$$

The l-moment ratios are l-coefficient of variation (l-Cv; $\tau_2 = \lambda_2/\lambda_1$), l-skewness (l-Cs; $\tau_3 = \lambda_3/\lambda_2$) and l-kurtosis (l-Ck; $\tau_4 = \lambda_4/\lambda_2$), respectively.

2.3.2 Regionalization

Researchers have focused on spatial variability of hydrological response of a given region to delimitate homogeneous hydrological regions called as hydrological regionalization. The definition of a homogeneous hydrological region is that the sites in that region show spatially a high degree of similarity from the hydrological response point of view. Thus, the limited information available at a site is able to be augmented and enhanced with information available at other sites in

the homogeneous region. Therefore, many approaches related to regionalization were developed, in the recently, the most popular of them is regionalization based on l-moments. The regionalization procedure used in this study is outlined below.

2.3.3 Discordancy measure

Main objective of this analysis is to identify any site in the selected region in three-dimensional space. The discordancy measure *Di* (Hosking and Wallis, 1997) compares the L-moment ratios of a site with those of the pooling group as a whole. If a given site is not in the cloud of (τ_3, τ_4) points on the l-moment diagram, that is, is far from the center of the cluster, the site is removed to other region. The sites in the homogeneous region (pooling group) form a cluster. Discordancy measure (D_i) of a site can be calculated by

$$\bar{u} = N^{-1} \sum_{i=1}^{N} u_i \tag{7}$$

$$S = (N-1)^{-1} \sum_{i=1}^{N} (u_i - \bar{u})(u_i - \bar{u})^T \tag{8}$$

$$D_i = \frac{1}{3} (u_i - \bar{u})^T S^{-1} (u_i - \bar{u}) \tag{9}$$

Where u is the vector of l-moments, and N is the number of stations. For $N^3 15$, Di should be less than or equal to 3.

2.3.4 Heterogeneity analysis

Hosking and Wallis (1993) recommended heterogeneity (H) test to assess whether the regions proposed as homogeneous according to discordancy measure of site characteristics are reasonably treated as a homogeneous region. This test compares the inter-site variation (dispersion) in sample l-moments for the group of sites. The homogeneity measures (H) are based on the simulation of 500 homogeneous regions with population parameters equal to the regional average sample l-moment ratios (Hosking and Wallis, 1997; Tallaksen, 2004).

This test for homogeneous of a region is based on

$$H = (V_{obs} - \mu_v)/\sigma_v \tag{7}$$

$$V = \left\{ \sum_{i=1}^{N} n_i (\tau_2^i - \tau_2^R)^2 / \sum_{i=1}^{N} n_i \right\}^{1/2} \tag{8}$$

Where V is the weighted standard deviation of the at-site sample L-CVs (t), and μ_V and σ_V are the mean and standard deviation of V, found through simulation. The simulation is performed by

fitting a Kappa distribution to the regional average L-moment ratios, 1, τ_2^i , τ_3^i and τ_4^i. The n_i is record length at site i, τ_2^i are the sample l-coefficient of variation (LCv), respectively. The value of the H-statistic indicates that the region under consideration is acceptably homogeneous when $H < 1$, possibly heterogeneous when $1 \leq H < 2$, and definitely heterogeneous when $H \leq 2$.

2.3.5 Choosing the regional frequency distribution

The regional frequency distribution is chosen based on the goodness-of-fit-test, Z^{DIST}, (Tallaksen, 2004). The statistics are given as:

$$Z^{DIST} = (\tau_4^{DIST} - \bar{\tau}_4 + ß_4)/\sigma_4 \tag{14}$$

$$ß_4 = N_{sim}^{-1} \sum_{m=1}^{N_{sim}} (\bar{\tau}_{4m} - \bar{\tau}_4) \tag{15}$$

$$\sigma_i = \left\{ (N_{sim} - 1)^{-1} \sum_{m=1}^{N_{sim}} (\tau_{4m} - \tau_4)^2 - N_{sim} ß_4^2 \right\}^{1/2} \tag{16}$$

where DIST is the candidate statistical distribution, τ_4^{DIST} is the population l-kurtosis of selected distribution, $\bar{\tau}_4$ is the regional average sample l-kurtosis, $ß_4$ is the bias of regional average sample l-kurtosis, σ_4 is the standard deviation of regional average sample l-kurtosis, and N_{sim} is realization of a region with N sites. Hosking and Wallis (1997) imply that the four parameter Kappa distribution for simulations includes a special case of the generalized logistic, generalized extreme values and generalized Pareto distributions, therefore, this distribution has capability of representing many of distribution. They judged from simulations that the value of 500 for N_{sim} should usually be adequate. Therefore, $ß_4$ and σ_4 parameters were estimated by using he four parameter Kappa distribution simulating 500 regions similar to the actual region. The parameters belonging to Kappa distribution were estimated by using the regional average l-moment ratios. A reasonable criterion being $|Z^{DIST}| \leq 1.64$ for an appropriate regional distribution, but the distribution giving the minimum $|Z^{DIST}|$ is considered as the best-fit distribution for the region.

The regional frequency analysis of seasonal and annual rainfall depths over Cekerek watershed was achieved by using the Fortran routines developed by Hosking (1996).

2.4 Data mining

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with a great potential to help decision-makers focus on the most important information in their data warehouses. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources, and can be integrated with new products and systems.

Data mining appears under a multitude of names, which includes knowledge discovery in databases, data or information harvesting, data archaeology, functional dependency analysis,

knowledge extraction, and data pattern analysis. In addition, there exist a large number of definitions for this group of methods. The term data mining is used for both the whole process of knowledge discovery and also for the specific algorithms which are used to achieve this aim. Among the several definitions of data mining, the most appropriate for real-world applications is given by Fayyad (1996): Data mining is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. In other words, data mining is the search for relationships and global patterns that exist among parameters, but are hidden among the data. The data mining technique used in this study for detecting drought categories with rules related to monthly rainfalls over Cekerek Watershed is the induction tree technique (See5), as described in Quinlan (1997).

2.4.1 The See5 algorithm

Learning systems based on decision trees are the easiest to use and to understand of all machine learning methods. Moreover, the condition and ramification structure of a decision tree is suitable for classification problems. The successive branches of a decision tree achieve a series of exhaustive and exclusive partitions among the set of objects that a decision maker wants to classify. The See5 algorithm is the latest version of the ID3 and See5 algorithms developed by Quinlan (1997). The criterion employed in See5 algorithm to carry out the partitions is based on some concepts from Information. Theory has been improved significantly over time. The main idea shared with similar algorithms is to choose the variable that provides more information based on entropy theory to realize the appropriate partition in each branch in order to classify the training set. The entropy is a measure of the randomness or uncertainty of a variable or a measure of the average amount of information that is supplied by the knowledge of a variable. The See5 algorithm uses entropy criteria in the separation of branches and nodes of the tree. A separation criterion for each node "t " is determined by using the equation:

$$Entropy = \sum_{i} - p_i \log p_i \tag{17}$$

Where the probability of the p^{th} cluster is located in node t. This quality and separation are carried out for the case of minimum entropy. In this case, a See5 significance test is carried out over the entire tree between the main nodes and children. As a result of this test, the child is pruned if the difference between the children and the mother is not significant (Sudha, 2006). Furthermore, See5 algorithm includes additional functions such as a method to change the obtained tree into a set of classification rules that are generally easier to understand than the tree. The See5 algorithm developed by Quinlan (1997) is the estimation of the class of a state over the amount of the other characteristics. See5 can correct decision trees by a classifying function or decision rules. Every rule in the program includes statistics with the rule number. Statistics (n, lift x) or (n/m, lift x) summarize the performance of the rule. Similar to a leaf, n represents the number of states coinciding with the rule during the training (correct estimation) and m represents number of states not placed in the class estimated by the rule (incorrect estimation). Accuracy of the rules is estimated by the Laplace rate (n – m +1) / (n + 2), and lift x = ((n – m + 1)/(n + 2))/(number of states for each class/total number of states). The Laplace accuracy rate is the most significant and useful statistic in the evaluation of the rules (Quinlan, 1997).

The See5 algorithm can also handle missing data and when a value is not known at a node

of the decision tree, it explores all possible outcomes and combines the resulting classifications arithmetically and chooses the class with the highest probability as ''the'' predicted class. A decision tree then also represents a set of control rules, with the characteristic that the rule set is structured such that only one rule is activated for any given and complete case. There are methods to generate more general rule sets from decision trees, but for the simulations here only complete decision trees were used. One of the advantages of decision trees as data mining algorithms is that such a set of rules can be derived, and the validity of these rules can be tested against other examples and domain experts can decide on the quality of the rules. This stands contrary to other data mining methods, such as neural networks, which act as a ''black box'' and it cannot be derived how the prediction is achieved there (Florian, 2003). In a decision tree, data are compiled and rules are written in an "if-conditional" style by moving from the roots of the tree to the leaves. Driving the rules in this way provides confirmation of the data mining results. These rules may be then shown to an expert and inspected as to whether or not the results are meaningful in practice (Solomatine and Dulal, 2003).

Data is divided into two parts in data mining model creation. The first part is used for training and the second part is used for testing. Data training can be performed in a supervised or unsupervised fashion. In supervised training (classification) labels indicating the classes of observations are attached to trained data (observations, measurements etc.) and then new data are classified based on trained data sets. In unsupervised training (clustering), class labels of training data are not known. Class labels in observed and measured data sets are determined by using current classes or clusters (Han and Kamber, 2006).

3. Results and discussion

The results belonging to seasonal and annual regional drought analysis based on SPI and decision tree technique, which is a data-mining approach, were given in sequence with the following.

3.1 The results based on l-moment approach

In order to achieve regional frequency analysis of the seasonal (SRS) and annual (ARS) rainfall depth series from the rainfall gauge stations over Cekerek watershed, some basic l-moment statistics, which are l-coefficient of variation (l-Cv), l-skevness (l-Cs) and l-kurtosis (l-Ck), were calculated for each station and given in Tables II, III and IV. Hosking (1990) implied that l-moment ratios of a series were bounded with $0 < \tau_2 < 1$, $-1 < \tau_3 < 1$ and $\frac{1}{4}(5\tau_3^2 - 1) \leq \tau_4 < 1$ for the l-Cv (τ_2), l-Cs (τ_3) and l-Ck (τ_4), respectively. As it can be seen in Tables II through V, these conditions were satisfied for the rainfall series. For the purpose of regionalization, it is important to check the existence of discordant station and homogeneity of the region. The first step was to apply discordancy and homogeneity tests to the data sets from the study region to judge whether all rainfall gauge stations of whole region formed a group of homogeneous sites or not. The discordancy test results related to the SRS data for k1 reference period (Jan-March) showed that the study area could not be taken into consideration as a homogeneous region due to existence of sites with $D_i > 3$. The region was divided into two sub-regions for k1 reference period (Jan-March), and there was no discordant site for each sub-region. Tables II-IV show discordancy measures (D_i) concerning with the SRS

Table III. Discordancy analysis results of rainfall gauge stations over the k3 and k4 reference period.

Stations	k3 reference period (July-September)				k4 reference period (October-December)			
	l-Cv	l-Cs	l-Cs	Di	l-Cv	l-Cs	l-Ck	Di
Corum	0.3379	0.2402	0.1722	1.02	0.2087	0.1216	0.1032	0.30
Eymir	0.4713	0.1321	–0.1099	1.84	0.1989	–0.0355	–0.0673	1.22
Ortakoy	0.3229	0.2149	–0.0475	1.53	0.2210	0.0253	0.1625	1.42
Alaca	0.4163	0.3248	0.1605	0.83	0.2259	0.0916	0.1617	0.35
Aydincik	0.1874	0.1743	0.0605	0.81	0.2422	0.4046	0.3940	1.83
Resadiye	0.3326	0.0786	0.0578	1.14	0.2129	0.2029	0.2436	0.57
Goynucek	0.3446	0.1214	0.0020	0.43	0.2020	0.2470	0.2737	0.51
Comarkoy	0.4495	0.2852	0.0527	1.77	0.1482	0.2240	0.1307	1.39
Mecitozu	0.2679	0.2203	0.1343	0.29	0.1570	0.1295	0.0626	0.72
Camlibel	0.3368	0.0220	–0.0009	1.75	0.1398	–0.0047	0.1752	2.27
Akdagmadeni	0.3239	0.2087	0.1203	0.10	0.2496	0.1447	0.1051	0.90
Ekinli	0.4115	0.3283	0.2091	0.55	0.2663	0.1790	0.1350	0.16
Sulusaray	0.3741	0.3663	0.2018	0.90	0.1918	0.1005	0.1172	0.11
Artova	0.2882	0.2928	0.0998	1.23	0.2200	0.1506	0.0764	0.34
Karamagara	0.2744	0.3349	0.2611	1.49	0.3483	0.3504	0.3187	1.31
Yoklaya	0.3381	0.2132	0.0974	0.11	0.2667	0.0448	–0.0722	1.44
Evciyenikisla	0.3576	0.2207	0.1544	1.21	0.2794	0.1855	0.1820	2.17

Table IV. Discordancy analysis results for annual (ARS) rainfall series.

Stations	Sample size (year)	Sample l-moments				Regional Statistics	
		l-Cv	l-Cs	l-Cs	Di		
Corum	75	0.0935	0.0399	0.0965	0.10		
Eymir	10	0.0913	0.0801	0.1907	0.33		
Ortakoy	17	0.0758	0.1423	0.1090	0.98		
Alaca	38	0.1133	–0.0045	0.0647	0.48		
Aydincik	15	0.1546	0.0244	0.2294	1.30	Mean	433.48
Resadiye	36	0.0967	0.1946	0.2101	1.78	l–Cv	0.1083
Goynucek	22	0.0920	0.0872	0.1090	0.12	l–Cs	0.0696
Comarkoy	13	0.0838	0.0336	0.0795	0.19	l–Ck	0.1270
Mecitozu	11	0.0512	–0.0662	0.2712	2.71	H1	–0.1379
Camlibel	11	0.0548	–0.0149	0.0859	0.90	H2	–0.1262
Akdagmadeni	26	0.1247	0.1407	0.2282	1.11	H3	–0.0613
Ekinli	33	0.1181	0.0935	0.1457	0.06	Z_{DIST}	GLO (0.04)
Sulusaray	31	0.0893	–0.1150	0.0851	1.81		
Artova	23	0.0924	0.0526	–0.0146	1.24		
Karamagara	31	0.1934	0.1609	0.2431	1.44		
Yoklaya	14	0.0999	0.1449	0.0451	1.01		
Evciyenikisla	32	0.1380	0.1314	0.0805	1.46		

and ARS data from the rainfall gauge stations in the region formed. The tables present that the discordancy measures for k-reference periods are smaller than 3 for each site. This emphasizes that there is no the discordant station in the region, and the sites in the region form a cluster.

Table V. Regional statistics for seasonal (SRS) rainfall series.

Regional Statistics	k-reference periods				
	k11	k12	k2	k3	k4
Mean	104.58	124.55	161.91	40.81	122.27
l-Cv	0.1893	0.2207	0.1978	0.3471	0.2301
l-Cs	–0.0002	0.0912	0.0825	0.2380	0.1585
l-Ck	0.0312	0.1045	0.1736	0.1252	0.1551
H1	–0.1137	–0.1543	–0.1273	–0.1910	–0.1747
H2	–0.1027	–0.1438	–0.1365	–0.2023	–0.1944
H3	–1.8947	–0.8516	0.8106	–0.2061	–0.0014
Z_{DIST}	GPA (–1.34)	GEV (0.95)	GLO (–0.26)	PIII (0.85)	GEV (–0.57)

For the k-reference periods, the homogeneity measures called as H_1, H_2 and H_3 based on l-Cv, l-Cs and l-Ck were smaller than one, except for H3 (–1.8947) belonging to k11 reference period (Jan-March) (See Table IV and Table V). The results stress that the regions formed for the k-reference periods can be considered as homogeneous. But, it is noted that the k11 reference period is possibly heterogeneous according to H3, due to $1 \leq |1.8947| < 2$. Whereas, the k11 reference period is acceptably homogeneous according to H1 and H2, owing to H1(–0.1137) and H2(–0.1027) < 1. In the regional frequency analysis studies, the H1 measure based on l-Cv is commonly used as the H1 heterogeneity measure has more discriminatory power to discriminate between homogeneous and heterogeneous regions. In this study, H1 measure was taken into consideration as a key indicator in forming homogeneous regions. In fact, Hosking and Wallis (1997) stated that the measures (H2 and H3) based on combination of l-Cv and l-Cs, and combination of l-Cs and l-Ck rarely yielded H values bigger than 2.

3.2 Goodness-of-fit-test

The goodness of fit test measure $|Z^{DIST}|$ was calculated for five distributions, namely, Generalized Logistic (GLO), Generalized Extreme Values (GEV), Generalized Normal or 3-parameter Log Normal (LN3), Pearson Type III (PE3) and Generalized Pareto (GPA) distributions, which are commonly used in hydrological studies. Among these distributions, the distribution with the smallest value, $|Z^{DIST}| \leq 1.64$, for the k-reference period was selected as the regional distribution. The GPA (1.34) for k11 reference period, the GEV (0.95), LN3 (1.18) and PE3 (0.99) for k12 reference period, the GLO (0.26) for k2 reference period, the PE3 (0.85) and GPA (1.30) for k3 reference period, the GEV (0.57), LN3 (0.73), PE3 (1.32) and GLO (1.51) for k4 reference period and the GEV (0.41), LN3 (0.04) and PE3 (0.12) for k5 reference period (annual) were estimated, respectively. All of these results express that the selected distributions for the k-reference periods can be used as a regional distribution, since the absolute values of estimated Z scores for the distributions were within the given criteria, $|Z^{DIST}| \leq 1.64$. The candidate regional distributions for k-reference periods were the GPA, GEV, GLO, PE3, GEV and LN3 with the smallest $|Z^{DIST}|$ value, respectively.

3.3 The regional SPI results

The regional SPI results for the homogeneous regions related to k-reference periods were give in Table VI. This table shows that the SPI values estimated by using the candidate regional distribution

Table VI. The regional total SPI results of drought categories for the homogeneous regions.

Drought category	k-reference periods					
	k11	k12	k2	k3	k4	k5
EW	0	0	1	11	1	0
VW	0	2	0	14	3	2
MW	2	5	4	41	4	13
N	32	75	220	133	132	186
NN	103	130	216	137	287	225
MD	41	6	6	47	5	9
SD	25	0	0	13	0	2
ED	7	0	0	9	0	1

Table VII. The rules from the decision tree approach for k-reference periods.

k-reference periods	Rule-1	Rule-2	Rule-3	Rule-4	Rule-5
k11	(24/1, lift 6.9)	(89, lift 1.9)	(34, lift 4.9)	(20, lift 8.3)	(6, lift 25.2)
(Jan-March)	P > 138.2	84.4≤ P < 138.2	62.5≤ P< 84.4	50.8 ≤ P < 62.5	P ≤ 50.8
	Class N [0.923]	Class NN [0.989]	Class MD [0.972]	Class SD [0.955]	Class ED [0.875]
k12	6/1, lift 26.1)	(60, lift 2.9)	(106, lift 1.6)	(2, lift 65.3)	
(Jan-March)	P > 230	138.3 ≤ P < 230	35.2 ≤ P < 138.3	P ≤35.2	
	Class MW [0.750]	Class N [0.984]	Class NN [0.991]	Class MD [0.750]	
k2	(44/1, lift 2.2)	(55, lift 1.8)	(2, lift 37.9)		
(April-June)	P > 156.3	53.8 ≤ P < 156.3	P ≤ 53.8		
	Class N [0.957]	Class NN [0.982]	Class MD [0.750]		
k3	(31/9, lift 2.0)	(24, lift 2.6)	(9/7, lift 8.7)	(5, lift 11.0)	(2, lift 24.0)
(July-Sep)	P > 34	16.3≤ P < 34	10.2≤P <16.3	6.1 ≤ P< 10.2	P ≤6.1
	Class N [0.697]	Class NN [0.962]	Class MD [0.273]	Class SD [0.857]	Class ED [0.750]
k4	(25, lift 3.2)	(57, lift 1.4)	(2, lift 31.5)		
(Oct-Dec)	P > 131.6	39.2 ≤ P < 131.6	P ≤ 39.2		
	Class N [0.963]	Class NN [0.983]	Class MD [0.750]		
k5	(34/1, lift 2.7)	(57, lift 1.6)	(3/1, lift 28.2)		
(annual)	P > 427.5	264 ≤ P <427.5	P ≤ 264		
	Class N [0.944]	Class NN [0.983]	Class MD [0.600]		

P: Rainfall depth.

for the region formed as homogeneous were frequently in the N and NN drought categories. (see Table I for abbreviations) but, it is interesting that the SPIs for k3 reference period (July-Sep) were scattered in all of drought categories, although the k3 period is the most drought season. The EW and VW drought categories have significant numbers in the k3 period as well when compared with other reference periods. This implies that the heavy storms occur in summer.

As the reference period increases to k5 reference period (Jan-Dec), the SPI values respond more slowly to short-term precipitation variation and the cycles of positive and negative SPI values become more visible. When the k-reference period is small the SPI is frequently above and below zero. The SPI for longer k-reference periods changes slowly owing to changes in precipitation totals.

3.4 Drought prediction based on decision tree

In homogeneous regions of Cekerek Watershed, monthly rainfall depths were taken into consideration as main parameter to delineate drought based on decision tree using the SPI drought categories in Table I. Hence, the data sets of the training and testing phases for k-reference periods were constituted. The number of samples in data sets consisted of covering monthly rainfalls are: 173 and 37 for k11 (Jan-March), 174 and 44 for k12 (Jan-March), 357 and 90 for k2 (April-June), 320 and 85 for k3 (July-Sep), 340 and 95 for k4 (Oct-Dec), and 350 and 88 for k5 (Jan-Dec), for training and testing phases respectively. The Table VII presents the rules based on decision tree approach for the training phase. The different rule numbers were defined for each reference periods. Table VII illustrates the followings: the rule, number of occurrences, accuracies and lifts, e.g. for the rule number 1 of the k11 reference period is that if the rainfall depth is bigger than 138.2 mm, drought category is normal. This condition occurred 24 times in the training phase. The value of 0.923 shows the rule-1's accuracy estimated by Laplace ratio. The lift 6.9 is the result of dividing the rule's accuracy by the relative frequency of the predicted drought class. The other rules in the table can be commented in a similar way described above. The falsely and correctly classified the numbers of cases for each reference periods are given in Table VIII along with the error percentages.

Table VIII. Categories of the classified number of cases for each reference periods based on the decision tree technique.

		k-reference periods					
		k11	k12	k2	k3	k4	k5
Number of cases	Falsely classified*	2	2	3	16	5	3
	Correctly classified	35	42	90	69	95	88
	Error (%)	0.05	0.05	0.03	0.19	0.05	0.03

*Details of falsely classified classes were given in the text

The rules belonging to training phase for each k-reference periods were applied to the monthly cumulative rainfall data sets separated for testing phase. These results are given as: The five rules in the training phase for k11 reference period were applied to the 37 monthly rainfall data set separated for testing phase. The model incorrectly classified two out of 37 rainfalls. They were considered as in N and SD categories whereas, these rainfall amounts should have been actually in MW and ED categories, respectively. The extreme values in rainfall data causes presumably the false categorization of the two rainfalls. The remaining rainfall amounts (9, 14, 7 and 5) were in N, NN, MD and SD drought categories, respectively. The accuracy of prediction obtained by using the rules decided in the training phase for k11 period was 95 %.

The four rules assigned for the testing phase of k12 period were applied to the 44-rainfall data. The two of the 44-rainfall data were in MW and NN drought categories instead of VW and MD. The other rainfall amounts (15, 24 and 3) were in N, NN and MD drought categories, respectively. The prediction accuracy for four rules was estimated as 95.5 %.

The amount of monthly rainfalls used in the testing phase of the k2 reference period (April-June) was 90. One rainfall of EW case and one rainfall of MW case and one rainfall of NN case were mistakenly categorized as N. The remaining monthly rainfalls (47, 37 and 3) were in N, NN

and MD drought categories, respectively. The prediction accuracy of the three rules designated for the k2 reference period was 96.7 %. The existence of extreme values in the time series of rainfall data of the testing phase may cause the three rainfalls be incorrectly classified.

The 85-monthly rainfall data in testing phase for k3 reference period (July-Sep) were taken into consideration. The 16- monthly rainfalls were classified as N and SD instead of EW, VW, MW and ED, respectively. The remaining rainfalls were in N, NN, MD and SD drought categories. The prediction accuracy of the five rules formed for the k3 reference period was 81.2%. The main reason in declining of the prediction accuracy is mostly the fluctuation in the data separated for the training and testing phases.

The five of the 95-monthly rainfalls used in the testing phase of the k4 reference period (Oct-Dec) were classified as N instead of EW, VW, MW and NN, respectively. The remaining monthly rainfalls (35, 53 and 2) were in N, NN and MD drought categories, respectively. The prediction accuracy of the three rules defined for the k4 reference period was 94.7%. The reason of the five rainfalls were incorrectly classified seems to be the existence of extreme values in the time series of rainfall data of the testing phase.

The three rules formed in the training phase for k5 (annual) reference period were applied to the 88-monthly rainfall data set for the testing phase. The twelve of the 88-rainfalls were classified as N, NN and MD instead of VW, MW, MD, SD and ED, respectively. The remaining rainfall amounts (38, 36 and 2) were in N, NN and MD drought categories, respectively. The accuracy of prediction for the tree rules was 86.4%.

In general, the difference among the monthly rainfall amounts separated for training and testing phase reduced the prediction accuracy of the model based on decision tree. The results showed that the decision tree approach was a good tool to predict drought occurrences. As described above, the prediction accuracy of the approach is considerably high. The comparison related to the number of rainfalls fallen in drought categories calculated from the general SPI algorithm and the decision tree technique (DT) was presented in Figures 2 through 7 for each k-reference periods. The figures also imply that the monthly cumulative rainfalls formed for the related k-reference periods in Cekerek Watershed are commonly in “Normal” and “Near Normal” drought categories.

4. Conclusions

In this study, it was aimed to perform seasonal regional drought analysis based on standardized precipitation index (SPI) and decision tree technique and results of both methods were compared. For this reason, the cumulative seasonal and annual rainfall series (SRS and ARS) for the k-th reference periods by using monthly rainfalls from 17 stations in Cekerek Watershed were constituted. The regionalization has been implemented by the method of l-moments. Two homogeneous region were formed for the k1 reference period, the watershed is taken into account as a whole for the other reference periods. Based on the goodness of fit test measure $|Z^{DIST}|$ the candidate regional distributions for k-reference periods (k11, k12, k2, k3, k4 and k5) (k1 (Jan-March), k2 (April-June), k3 (July-Sep), k4 (Oct-Dec) and k5 (annual) time periods), were the GPA, Generalized Extreme Values (GEV), Generalized Logistic (GLO), Pearson Type III (PE3), GEV and 3-parameter Log Normal (LN3), respectively. The SPI algorithm is used directly when a given data is normally distributed. Therefore, an equiprobability transformation was applied from the fitted regional distribution to the standard normal one.

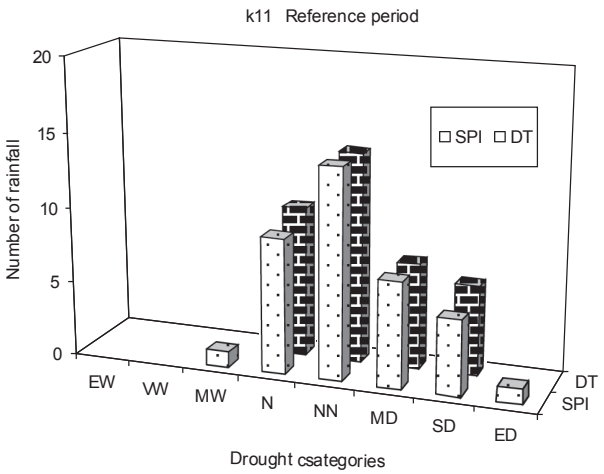

Fig. 2. Comparison related to the number of rainfalls from the SPI and DT algorithms.

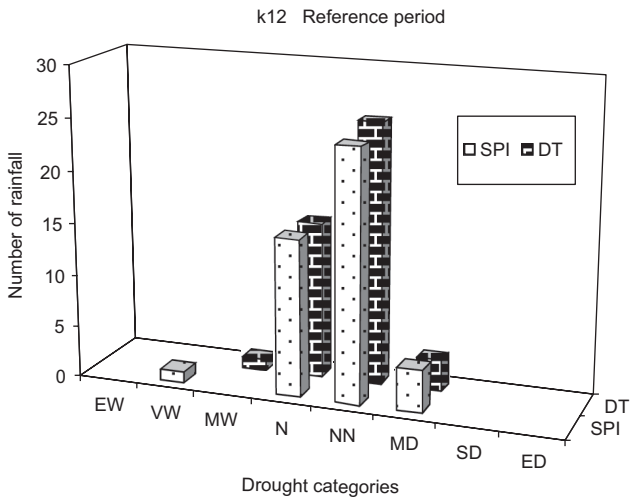

Fig. 3. Comparison related to the number of rainfalls from the SPI and DT algorithms.

The regional SPI results based on the candidate regional distributions present that the N and NN drought categories were frequently observed in all of the sub-regions. But, it is surprising that the cumulative monthly rainfall amounts for the most drought season in Cekerek watershed, the k3 reference period, were scattered in all of drought categories. The EW and VW drought categories were estimated in significant numbers in k3 period as well when compared with other periods. This may be the results of the occurrences of the heavy storms in summer. When the k-reference period is small, the SPI is frequently above and below zero value. The SPI for longer k-reference periods changes slowly owing to changes in precipitation.

The monthly cumulative rainfall data sets separated as the training and testing phases for k-reference periods were constituted. The drought categories for each k-reference period were predicted by applying the decision tree's rules obtained from the training phase to the rainfall

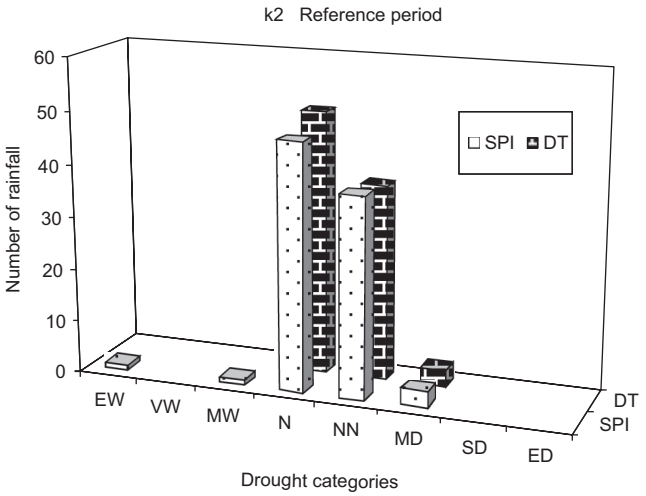

Fig. 4. Comparison related to the number of rainfalls from the SPI and DT algorithms.

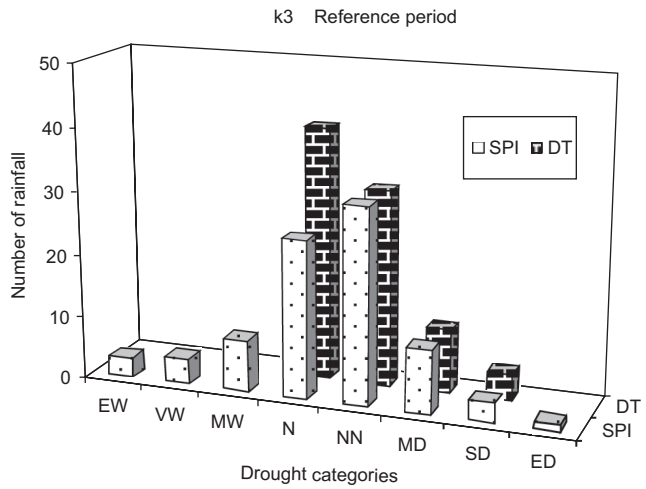

Fig. 5. Comparison related to the number of rainfalls from the SPI and DT algorithms.

data sets in testing phase. The results showed that there was no significance difference between drought categories from the conventional SPI algorithm and decision tree approaches. Moreover, the accuracy of prediction by decision tree approach for k-reference periods was greater than 94 %, except for k3 and k5 reference periods. The prediction accuracy of the k3 and k5 reference periods was 81.2 and 86.4 %, respectively. Understanding drought, which is a creeping phenomenon, is a very difficult task. Therefore, drought prediction is very important challenge for researcher, water resource planners, and local administrations. This paper will highly contributed to preventing ecosystem from the damage of drought occurrences. Quantifying the temporal patterns of drought based on the precipitation amount will help the policy makers to allocate water demands and to manage water resources especially during drought periods. This paper demonstrates the decision tree technique could serve to understand the current patterns of precipitation for such purposes.

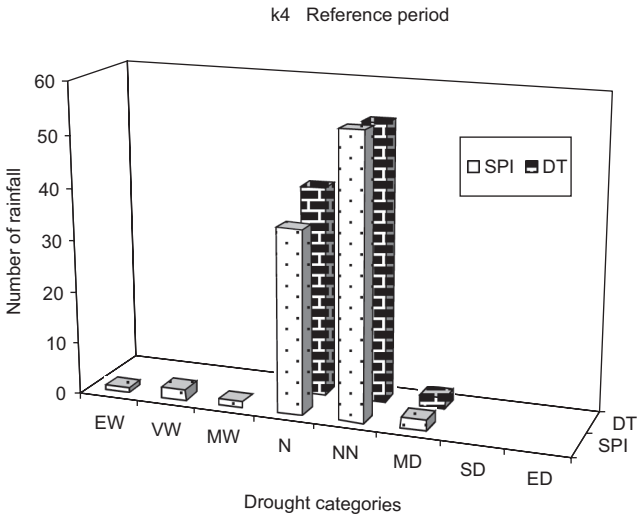

Fig. 6. Comparison related to the number of rainfalls from the SPI and DT algorithms.

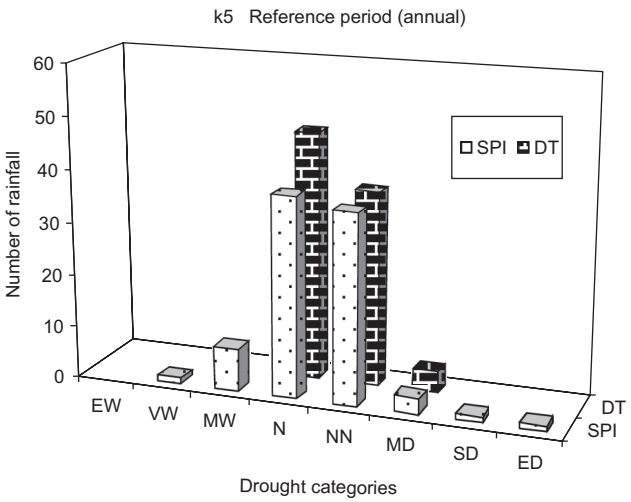

Fig. 7. Comparison related to the number of rainfalls from the SPI and DT algorithms.

References

Agnew C. T., 1990. Spatial aspects of drought in the Sahel. *J. Arid Environ*. **18**, 279-293.

Agnew C. T. and A. Warren, 1996. A framework for tackling drought and degradation. *J. Arid Environ*. **33**, 309-320.

Alatise M. O. and O. B. Ikumawoyi, 2007. Evaluation of drought from rainfall data for lokoja. A confluence of two major rivers. Electronic Journal of Polish Agricultural Universities Tomo 10, 1, Art. 5, Ondo State, Nigeria Available Online: http://www.ejpau.media.pl/volume10/issue1/art-05.html. Accessed in February 2010.

Alley W. M., 1984. The palmer drought severity index: limits and assumptions. *J. Clim. App. Clim. Meteorol.* **23**, 1100-1109.

Anonymous, 1970. Soils of Yeşilirmak Basin. General Directorate of Soil and Water Publications, Ankara.

Belda F. and M. C. Penadés, 2007. Using data-mining techniques for monitoring climatic variations. Application to drought. 7th EMS Annual Meeting / 8th European Conference on Applications of Meteorology, San Lorenzo de El Escorial, Spain, 01-05 October. Available online: http://meetings.copernicus.org/www.cosis.net/abstracts/EMS2007/00290/EMS2007-J-00290.pdf. Accessed in : February 2010

Edwards D. C. and T. B. McKee, 1997. Characteristics of 20th century drought in the United States at multiple time scales. Climatology Report Number 97-2, Colorado State University, Fort Collins, CO.

Fayyad U. M., G. Piatetsky-Shapiro and P. Smyth, 1996. From data mining to knowledge discovery: An Overview. In: Advances in knowledge discovery and data mining (U. M. Fayyad, Ed.). AAAI Press and MIT Press, USA, pp 1-34.

Florian T. B., A. S. Dragan and A. W. Godfrey, 2003. Water reservoir control with data mining. *J. Water Res. Manage.* **129**, 26-34.

Greenwood J. A., J. M. Landwehr, N. C. Matalas and J. R. Wallis, 1979. Probability weighted moments: Definition and relation to parameters of several distributions expressible in inverse form. *Water Resour. Res*. **15**, 1049-1054.

Giddings L., Soto M., Rutherford and M. B.M., Maarouf, 2005. Standardized precipitation index zones for Mexico. *Atmósfera* **18**, 33-56.

Guttman N. B., 1998. Comparing the Palmer drought index and the standardized precipitation index. *J. Am. Water Resour. As.* **34**, 113-121.

Guttman N. B., 1999. Accepting the Standardized Precipitation Index: A Calculation algorithm. *J. Am. Water Resour. As.* **35**, 311-322.

Han J. and M. Kamber, 2006. *Data mining: Concepts and techniques*. Morgan Kaufmann Publishers, New York, 664 pp.

Hosking J. R. M., 1990. L-Moments: Analysis and estimation of distributions using linear combinations of order statistics. *J. Roy. Stat. Soc. B*. **52**, 105-124.

Hosking J. R. M. and J. R. Wallis, 1993. Some statistics useful in regional frequency analysis. *Water Resour. Res* **29**, 271-281.

Hosking J. R. M., 1996. Fortran routines for use with the method of L-moments. Research Report RC 20525, Version 3, New York, USA, 33 pp.

Hosking J. R. M. and J. R. T. Wallis, 1997. *Regional frequency analysis: An approach based on L-moments*. Cambridge University Press, 34 pp.

Komuscu A. U., 1999. Using the SPI to analyze spatial and temporal patterns of drought in Turkey. Drought Network News **11**, 7-13.

Kumar M. N., C. S. Murthy, M. V. R. Sesha Sai and P. S. Roy, 2009. On the use of Standardized Precipitation Index (SPI) for drought intensity assessment. *Meteorol. Appl*. **16**, 381-389.

Labedzki L., 2007. Estimation of local drought frequency in central Poland using the standardized precipitation index SPI. *Irrig. Drain*. **56**, 67-77.

Lana X., C. Serra and A. Burguena, 2001. Patterns of monthly rainfall shortage and excess in terms of the standardized precipitation index for Catalonia (NE Spain). *Int. J. Climatology*. **21**, 1669-1691.

Le Houerou H. N., 1996. Climate change, drought and desertification. *J. Arid Environments* **34**, 133-185.

Loukas A. and L. Vasiliades, 2004. Probabilistic analysis of drought spatiotemporal characteristics in Thessaly region, Greece. *Nat. Hazards Earth Syst*. **4**, 719-731.

McKee T. B., N. J. Doesken and J. Kleist, 1993. The relationship of drought frequency and duration to time scales. 8th Conference on Applied Climatology, American Meteorological Society, Anaheim, California, 17-22 January, American Meteorological Society, Dallas, Texas, 15-20 January, Anaheim, CA. American Meteorological Society, Boston, MA, 179-184

McKee T. B., N. J. Doesken and J. Kleist, 1995. Drought monitoring with multiple time scales. Preprints, 9th American Meteorological Society, Dallas, TX, 233-236.

Oladipo E. O., 1985. A comparative performance analysis of three meteorological drought indices. *Int. J. Climatol.* **5**, 655-664.

Palmer W. C., 1965. Meteorological drought. Research Paper No. 45, U.S. Weather Bureau, Washington, DC, 58 pp.

Quinlan J. R., 1997. See5 (available from http://www.rulequest.com/see5-info.html) Accessed in February 2010.

Redmond K. T., 2000. Integrated climate monitoring for drought detection. In Drought: A Global Assessment (Wilhite D. A., Ed.), Hazards and Disasters Series, Routledge, London, 145-158.

Seiler R. A., M. Hayes and L. Bressan, 2002. Using the standardized precipitation index for flood risk monitoring. *Int. J. Climatol.* **22**, 1365-1376.

Sharma A., 2006. Spatial data mining for drought monitoring: An approach using temporal NDVI and rainfall relationship. Master thesis, The International Institute for Geo-information Science and Earth Observation, The Netherlands.

Solomantine D. P. and K. N. Dulal, 2003. Model trees as an alternative to neural networks in rainfall-runoff modeling. *Hydrolog. Sci. J.* **48**, 455-472.

Sudha V., N. K. Ambujam and K. Venugopal, 2006. A data mining approach for deriving irrigation reservoir operating rules. Conference on Water Observation and Information System for Decision Support, Orhid, Macedonia. Available: http://balwois.com/balwois/administration/full_paper/ffp-http://balwois.com/balwois/administration/full_paper/ffp-643.pdf. Accessed in February 2010.

Tadesse T., D. A. Wilhite S. K., Harms, M. J. Hayes and S. Goddard, 2004. Drought monitoring using data mining techniques: A case study for Nebraska, USA. *Nat. Hazards* **33**, 137-159.

Tallaksen L. M., H. Madsen and H. Hisdal, 2004. Frequency analysis. In: *Hydrological drought. Processes and estimation methods for streamflow and groundwater* (Tallaksen L. M., Van Lanen H. A. J. Eds.). Developments in water science, Elsevier Science B.V., Amsterdam, 199-271.

Thorn H. C. S., 1966. Some methods of climatological analysis. WMO technics/note number No. 81, 16-22.

Wu H., Hubbard, K. G. and D. A. Wilhite, 2004. An agrıcultural drought rısk-assessment model for corn and soybeans. *Int. J. Climatol.* **24**, 723-741.

Yamoah C. F., Walters D. T., Shapiro C. A., Francis C. A. AND M. J. Hayes, 2000. Standardized precipitation index and nitrogen rate effects on crop yields and risk distribution in maize. *Agr. Ecosyst. Environ.* **80**, 113-120.