Journal of Hydrology 550 (2017) 92–102

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Research papers

Regional frequency analysis using Growing Neural Gas network

Amin Abdi [a,*], Yousef Hassanzadeh [a], Taha B.M.J. Ouarda [b,c]

[a] Department of Civil Engineering, University of Tabriz, Tabriz, Iran
[b] Institute Center for Water and Environment, Masdar Institute of Science and Technology, Abu Dhabi, United Arab Emirates
[c] National Institute for Scientific Research, INRS-ETE, Quebec (QC), Canada

ARTICLE INFO

Article history:
Received 9 October 2016
Received in revised form 21 February 2017
Accepted 20 April 2017
Available online 23 April 2017
This manuscript was handled by A. Bardossy, Editor-in-Chief, with the assistance of Fi-John Chang, Associate Editor

Keywords:
Regional frequency analysis
Growing Neural Gas
Minimum Description Length
Clustering method
L-moments

ABSTRACT

The delineation of hydrologically homogeneous regions is an important issue in regional hydrological frequency analysis. In the present study, an application of the Growing Neural Gas (GNG) network for hydrological data clustering is presented. The GNG is an incremental and unsupervised neural network, which is able to adapt its structure during the training procedure without using a prior knowledge of the size and shape of the network. In the GNG algorithm, the Minimum Description Length (MDL) measure as the cluster validity index is utilized for determining the optimal number of clusters (sub-regions). The capability of the proposed algorithm is illustrated by regionalizing drought severities for 40 synoptic weather stations in Iran. To fulfill this aim, first a clustering method is applied to form the sub-regions and then a heterogeneity measure is used to test the degree of heterogeneity of the delineated sub-regions. According to the MDL measure and considering two different indices namely CS and Davies–Bouldin (DB) in the GNG network, the entire study area is subdivided in two sub-regions located in the eastern and western sides of Iran. In order to evaluate the performance of the GNG algorithm, a number of other commonly used clustering methods, like K-means, fuzzy C-means, self-organizing map and Ward method are utilized in this study. The results of the heterogeneity measure based on the L-moments approach reveal that only the GNG algorithm successfully yields homogeneous sub-regions in comparison to the other methods.

1. Introduction

Regional frequency analysis (RFA) is commonly utilized in hydrology to circumvent the limitations of at-site statistical estimation procedures due for instance to the unavailability or the short length of the data series (Ouarda et al., 2001; Zhang et al., 2012). The information obtained based on RFA is more valuable, flexible and accurate than the single-site analysis (Atiem and Harmancioglu, 2006). RFA usually has two main steps: the delineation of hydrologically homogeneous regions and the estimation of hydrological variables within each region (Leclerc and Ouarda, 2007; Charron and Ouarda, 2015; Wazneh et al., 2015; Abdi et al., 2016b,c). In the first step, the most complex and important one, the regions can be formed based on a clustering method and then tested by a heterogeneity measure (Abida and Ellouze, 2006; Ouarda et al., 2008; Basu and Srinivas, 2014; García-Marín et al., 2015).

Clustering algorithms are used to assemble objects into a set of specific groups with a maximum similarity between the members (Modarres, 2010). A number of clustering techniques are available, among which the most popular are the principal component analysis (PCA) (Iyengar and Basak, 1994; Singh and Singh, 1996; Chiang et al., 2002), Ward (Modarres, 2006; Kahya et al., 2008; Yang et al., 2010), K-means (KM) (Ngongondo et al., 2011; Dikbas et al., 2013; Rahman et al., 2013; Kulkarni, 2016), fuzzy C-means (FCM) (Rao and Srinivas, 2006a; Dikbas et al., 2012; Kar et al., 2012; Aydogdu and Firat, 2015) and self-organizing map (SOM) (Lin and Chen, 2006; Razavi and Coulibaly, 2013). In addition, various methods can be obtained by using PCA in association with a clustering method [e.g., Ward (Dinpashoh et al., 2004; Awadallah and Yousry, 2012), KM (Satyanarayana and Srinivas, 2008), FCM (Shamshirband et al., 2015; Asong et al., 2015) and SOM (Chen et al., 2011)]. There is no agreement between researchers about the superiority of any particular method. Most of the clustering algorithms have problems in dealing with high-dimensional data sets and determining non-spherical shapes of clusters (Steinbach et al., 2003). Because of the arbitrary shapes of regions and the effects of various watershed related attributes, which are inevita-

* Corresponding author.
E-mail addresses: amin.abdi@yahoo.com, aabdi@tabrizu.ac.ir (A. Abdi).

http://dx.doi.org/10.1016/j.jhydrol.2017.04.047
0022-1694/

ble in hydrological regionalization, selecting the best method is important (Basu and Srinivas, 2014).

In this paper, we present an application of the Growing Neural Gas (GNG) network for hydrological data clustering. The GNG algorithm, which is based on unsupervised artificial neural networks, was first introduced by Fritzke (1995). The GNG network is a clustering algorithm that works incrementally, i.e., the number of neurons will increase during the training procedure without using a prior knowledge concerning the structure of the input patterns (Oliveira Martins et al., 2009; Angelopoulou et al., 2015; Fink et al., 2015). Unlike classical clustering algorithms, the GNG algorithm has an adaptable network structure that makes it suitable for the task of learning the topology of high-dimensional data sets (Zaki and Yin, 2008; Linda and Manic, 2009; Bouguelia et al., 2015). This algorithm has gained significant interest in a number of fields, especially in the field of computer vision such as: image compression (García-Rodríguez et al., 2007); human gestures recognition (Angelopoulou et al., 2011; Botzheim and Kubota, 2012; García-Rodríguez et al., 2012), three-dimensional feature extraction (Donatti and Würtz, 2009; Viejo et al., 2012; Morell et al., 2014), and three-dimensional surface reconstruction (Noguera et al., 2008; Cretu et al., 2008; Rêgo et al., 2010; Fišer et al., 2013; Orts-Escolano et al., 2014; Jimeno-Morenilla et al., 2013, 2016). The GNG algorithm is also gaining increasing interest in a number of other fields such as medicine (Cselényi, 2005; Oliveira Martins et al., 2009; Angelopoulou et al., 2015); robotics (Carlevarino et al., 2000; Ferrer, 2014); economics (Lisboa et al., 2000; Decker, 2005); industrial applications (Cirrincione et al., 2011, 2012); communications (Bougrain and Alexandre, 1999), astronomy (Hocking et al., 2015); geography (Figueiredo et al., 2007); and biology (Ogura et al., 2003). To the authors' knowledge, there are still no studies that applied the GNG algorithm in the general fields of hydrology and water resources, and specifically to delineate homogeneous hydrological regions under the framework of RFA.

The quality of the formed clusters and the optimal number of clusters for a given data set can be determined by using the cluster validity indices (Rao and Srinivas, 2006b; Goyal and Gupta, 2014). For this purpose, the Minimum Description Length (MDL) principle, which has been widely applied in the field of neural networks, can be employed to evaluate the network's ability through balancing the capability and complexity of the network (Tenmoto et al., 1998; Bischof et al., 1999; Qin and Suganthan, 2004, 2005).

After the application of the GNG network, it is necessary to utilize a heterogeneity measure to determine the degree of heterogeneity of the delineated regions. In addition, the heterogeneity measure can offer a comparison between several clustering methods in order to find out which one yields regions that are more homogeneous (Basu and Srinivas, 2014). For this purpose, a number of heterogeneity measures have been proposed in the hydrologic literature. Among them, Hosking and Wallis (1993, 1997) proposed a measure based on the L-moments approach, which is known as the most powerful method in RFA (Viglione et al., 2007; Chebana and Ouarda, 2007; Ilorme and Griffis, 2013; Masselot et al., 2016). The L-moments method is widely used for the regional analysis of extreme hydrologic events such as droughts (Abolverdi and Khalili, 2010b; Núñez et al., 2011; Santos et al., 2011; Yoo et al., 2012), precipitations (Wallis et al., 2007; Satyanarayana and Srinivas, 2011; Hailegeorgis et al., 2013; Núñez et al., 2016), and floods (Srinivas et al., 2008; Gaume et al., 2010; Ilorme and Griffis, 2013; Nguyen et al., 2014).

In the present study, the GNG algorithm is applied for the RFA of drought severity in Iran during the period of 1971–2011. Then the results are compared by using the L-moments approach to those of a number of conventional algorithms, including Ward, KM, FCM and SOM. For this purpose, drought severity is extracted from the recently developed drought index called Multivariate Standardized Precipitation Index (MSPI), proposed by Bazrafshan et al. (2014, 2015). The MSPI, which is calculated based on the standardized precipitation index (SPI) and the principal component analysis (PCA), has the ability to aggregate the various time scales of the SPI into a new time series. In order to represent seasonal variations of precipitation throughout the year, a monthly time scale is considered for the MSPI.

2. Study area and data

In this study, the monthly precipitation data of 40 synoptic weather stations located in Iran were analyzed. The study area, Iran, covers an area of about 1,648,000 km^2, and lies between the latitudes 25°–40° North and longitudes 44°–64° East. The spatial distribution of the selected stations is fairly uniform across Iran. The selected stations have a record of 41 years covering the period from 1971 to 2011.

The two important mountain ranges of Iran are the Alborz and Zagros. Alborz, in located the northern part of Iran, extending along the southern Caspian Sea, while Zagros, is located in the western part of Iran, extending from the northwest to the southwest, impede Mediterranean moisture systems crossing through Iran (Shiau and Modarres, 2009). Most of the eastern part of the Iran is comprised of two great deserts called Dasht-e Kavir and Dasht-e Loot (Abolverdi and Khalili, 2010a). These mountain ranges and deserts have a great influence on the spatial and temporal distribution of precipitation and temperature over Iran (Dinpashoh et al., 2004).

The map of the study area and the spatial distribution of the stations are illustrated in Fig. 1. Also, Table 1 presents the following attributes of the stations: names, geographical variables and mean of the annual precipitations. The data set was supplied by the Meteorological Organization of Iran.

3. Methodology

In this section, the methodology proposed for drought RFA is described. For this purpose, first, the steps of the analysis procedure are presented, and then the proposed approaches are given in the following subsections. The steps of the procedure are as follows:

1. Consider a number of sites within the region, and assemble the monthly precipitation data for each site,
2. Calculate the MSPI values and extract drought severities,
3. Determine the sub-regions by using the clustering method and cluster validity index,
4. Compute the heterogeneity and discordancy measures using L-moments approach,
5. Test the sub-regions for regional homogeneity,
6. Adjust the heterogeneous sub-regions, and
7. Specify the homogeneous sub-regions.

3.1. Multivariate Standardized Precipitation Index (MSPI)

Bazrafshan et al. (2014) recently developed the MSPI index for drought monitoring. The MSPI is based on the several time series of the Standardized Precipitation Index (SPI) and the Principal Component Analysis (PCA) as a multivariate approach. Unlike the MSPI index, the SPI index does not have the flexibility to consider a variety of time scales. So, this may result in a confusion in the identification of drought periods. On the other hand, relating a cer-