Journal of Hydrology 509 (2014) 387–397

Contents lists available at ScienceDirect

Journal of Hydrology

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

Identification of homogeneous regions for regionalization of watersheds by two-level self-organizing feature maps

F. Farsadnia [a], M. Rostami Kamrood [b], A. Moghaddam Nia [c,*], R. Modarres [d], M.T. Bray [e], D. Han [f], J. Sadatinejad [g]

[a] Irrigation and Drainage, Ferdowsi University of Mashhad, Iran
[b] Irrigation and Drainage, Faculty of Agriculture, University of Zabol, Iran
[c] Faculty of Natural Resources, University of Tehran, Karaj, Iran
[d] INRS-ETE, University of Québec, 490 de la Couronne, Québec G1K 9A9, Canada
[e] Civil Engineering, Institute of Environment and Sustainability, Cardiff University, UK
[f] Civil Engineering, Faculty of Engineering, University of Bristol, Bristol, UK
[g] Department of Renewable Energies and Environment, Faculty of New Sciences and Technologies, University of Tehran, Iran

ARTICLE INFO

Article history:
Received 8 August 2011
Received in revised form 18 November 2013
Accepted 25 November 2013
Available online 4 December 2013
This manuscript was handled by Andras Bardossy, Editor-in-Chief, with the assistance of Vazken Andréassian, Associate Editor

Keywords:
Regionalization
Self-organization feature maps
Clustering methods
Hydrologic homogeneity
Cluster validation measures

SUMMARY

One of the several methods in estimating flood quantiles in ungauged or data-scarce watersheds is regional frequency analysis. Amongst the approaches to regional frequency analysis, different clustering techniques have been proposed to determine hydrologically homogeneous regions in the literature. Recently, Self-Organization feature Map (SOM), a modern hydroinformatic tool, has been applied in several studies for clustering watersheds. However, further studies are still needed with SOM on the interpretation of SOM output map for identifying hydrologically homogeneous regions. In this study, two-level SOM and three clustering methods (fuzzy *c*-mean, *K*-mean, and Ward's Agglomerative hierarchical clustering) are applied in an effort to identify hydrologically homogeneous regions in Mazandaran province watersheds in the north of Iran, and their results are compared with each other. Firstly the SOM is used to form a two-dimensional feature map. Next, the output nodes of the SOM are clustered by using unified distance matrix algorithm and three clustering methods to form regions for flood frequency analysis. The heterogeneity test indicates the four regions achieved by the two-level SOM and Ward approach after adjustments are sufficiently homogeneous. The results suggest that the combination of SOM and Ward is much better than the combination of either SOM and FCM or SOM and *K*-mean.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

In hydrology, estimating the frequency and magnitudes of extreme values such as floods, rainstorms and droughts is very important. Because extreme events are rare and their data records are often short, estimation of the frequencies of extreme events is difficult. Therefore regional frequency analysis is used for reliable estimation of hydrologic quantiles. In regional frequency analysis, a site must be assigned to a homogeneous region because an approximate homogeneity is required to ensure that a regional frequency analysis is more accurate than an at-site analysis (Hosking and Wallis, 1997).

When many sites are involved in a regional frequency analysis, identification of homogeneous regions is usually the most difficult part of the analysis and requires a great amount of subjective judgment. Cluster analysis has been used successfully to identify homogeneous regions for regional frequency analysis in hydrology. There are several methods for watershed clustering such as the *k*-means (Burn and Goel, 2000; Burn, 1989), agglomerative hierarchical clustering (Hosking and Wallis, 1997; Nathan and McMahon, 1990) and hybrid clustering (Rao and Srinivas, 2006).

The Self-Organizing feature Map (SOM) algorithm (Kohonen, 1982) is a heuristic model used to visualize and explore linear and nonlinear relationships in high-dimensional datasets. SOMs were firstly used in the 1980s in speech recognition (Kohonen et al., 1984). SOM is one of the widely used artificial neural networks (ANN) in many industrial applications such as pattern recognition, biological modeling, data compression, signal processing, and data mining (Kohonen, 2001). Recently SOM is used as a modern informatics tool to identify the hydrologically homogeneous regions. Hall and Minns (1999) used SOM in the classification of southwest England and Wales with five catchment characteristics per gauging site. They grouped the output neurons into three

* Corresponding author. Tel.: +98 9151441381; fax: +98 2632249313.
E-mail addresses: farhadfarsad@ymail.com (F. Farsadnia), a.moghaddamnia@ut.ac.ir (A. Moghaddam Nia), reza.modarres@ete.inrs.ca (R. Modarres), BrayM1@cardiff.ac.uk (M.T. Bray), D.Han@bristol.ac.uk (D. Han), jsadatinejad@ut.ac.ir (J. Sadatinejad).

0022-1694/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jhydrol.2013.11.050

distinct groups in order to obtain three homogeneous regions. Jingyi and Hall (2004) applied Ward's cluster Method, fuzzy *c*-means (FCM) method and SOM to classify the Gan and Ming River basin in south east of China. Their results indicate that SOM is preferable over the other two methods. Lin and Wang (2006) proposed a one-step method to perform cluster analysis and discrimination analysis based on SOM. They applied this method on the hydrological factors affecting low-flow duration curves in southern Taiwan. Lin and Chen (2006) applied the SOM, *K*-means method and Ward's method to actual rainfall data in Taiwan to identify homogeneous regions for regional frequency analysis. They used two-dimensional map to indicate eight clusters of rainfall. Their results showed that SOM could identify the homogeneous regions more accurately compared with the other two clustering methods. Herbst and Casper (2008) used SOM to obtain a topologically ordered classification and clustering of the temporal patterns present in the model outputs obtained from Monte-Carlo simulations. This clustering of the entire time series allowed them to differentiate the spectrum of the simulated time series with a high degree of discriminatory power and showed that the SOM could provide insights into parameter sensitivities, while helping to constrain the model parameter space to the region that best represents the measured time series. Ley et al. (2011) compared two kinds of inputs for SOM to investigate hydrological similarity of 53 gauged catchments in Rhinel and Palatinate, Germany. They compared groups of catchments clustered by response behavior with clusters of catchments based on catchment properties. Results show an overlap of 67% between these two pools of clustered catchments which can be improved using the topologic correctness of SOMs. Razavi and Coulibaly (2013) used five streamflow signatures to identify homogenous watersheds in the Province of Ontario, and compared them with the classified watersheds using the selected nonlinear clustering techniques including Self Organizing Maps (SOMs), standard Non-Linear Principal Component Analysis (NLPCA), and Compact Non-Linear Principal Component Analysis (Compact-NLPCA).

Although SOM has been successfully utilized as a first step in clustering algorithms, it is difficult to distinguish subsets because there are still no clear boundaries between possible clusters. Therefore, it is necessary to subdivide the map into different groups according to the similarity of the weight vectors of the neurons. In order to solve this problem, researchers tried several methods. Lampinen and Oja (1992) proposed a two-level SOM, where outputs of the first SOM are fed into a second SOM as inputs. This model performs better than SOM and classical *K*-means algorithms in classifying artificial data and sensory information from low-level feature detectors in a computer vision system. Vesanto and Alhoniemi (2000) applied both hierarchical agglomerative and partitional *K*-means clustering algorithms to group the output from SOM. They expressed that the most important benefit of this procedure was that computational load decreased considerably; moreover this method could cluster large data sets successfully as well as handle several different preprocessing strategies in a limited time. Srinivas et al. (2008) combined self-organizing feature map and fuzzy clustering to classify watersheds in Indiana, USA. They subdivided the region into seven homogeneous groups. Clearly, more research is still needed in this field to gain valuable experiences and explore alternative approaches.

In this study, we used three methods to divide the trained SOM units into several subgroups. First, the unified distance matrix algorithm (*U*-matrix) as a visual method was applied. Fuzzy *c*-mean algorithm, Ward's agglomerative hierarchical clustering (Ward) and *K*-mean methods were also applied to the trained SOM map to compare the subgroups separated by each method. In the next step, based on l-moment statistics the best method was selected. Finally adjusted regions are created by a two-level SOM and then the best regional distribution function and associated parameters are selected by the L-moment approach. Flow chart of the methodology proposed to determine hydrologically homogeneous regions is shown in Fig. 1.

2. Self-Organizing feature Map (SOM)

2.1. The SOM algorithm

SOM approximates the probability density function of input data through an unsupervised learning algorithm, and is not only an effective method for clustering, but also for the visualization and abstraction of complex data (Kohonen, 2001). The algorithm has properties of neighborhood preservation and local resolution of the input space proportional to the data distribution (Kohonen, 1982, 2001). A SOM consists of two layers: an input layer formed by a set of nodes (or neurons which are computational units), and an output layer (Kohonen layer) formed by nodes arranged in a two-dimensional grid (Fig. 2). The number of output neurons in an SOM (i.e. map size) is important to detect the deviation of the data. If the map size is too small, it might not explain some important differences that should be detected. Conversely, if the map size is too big, the differences are too small (Wilppu, 1997). The number of output neurons in an SOM can be selected using the heuristic rule suggested by Vesanto et al. (2000). The optimal number of map units is close to $5 \times \sqrt{N}$, where N is the number of samples in the data set.

Each node in the input layer is connected to all the nodes in the output layer by synaptic links. Each output node has a vector of coefficients associated with input data. The coefficient vector is referred to as a weight (or connection intensity) vector, W, between the input and output layers. The weights establish a link between the input units (i.e., feature vector) and their associated output units (i.e., groups of feature vector) (Fig. 2).

The algorithm can be described as follows: when an input feature vector X is presented to the SOM, the nodes in the output layer compete with each other, and the winning neuron (the neuron with the closest match to the presented input) is chosen. The winner and its neighbors, predefined in the algorithm, update their weight vectors according to the SOM learning rules as follows:

$$w_{ij}(t+1) = w_{ij} + \alpha(t) \cdot h_{jc}(t)[X_i(t) - w_{ij}(t)] \tag{1}$$

where $w_{ij}(t)$ is a weight between a node i in the input layer and a node j in the output layer at iteration time t, $\alpha(t)$ is a learning rate factor which is a decreasing function of the iteration time t, and $h_{jc}(t)$ is a neighborhood function (a smoothing kernel defined over the lattice points) that defines the size of neighborhood of the winning node (c) to be updated during the learning process. This learning process is continued until a stopping criterion is met, usually, when weight vectors stabilize or when a number of iterations are completed. This learning process results in the preservation of the connection intensities in the weight vectors. A detailed description of the SOM algorithm can be found in Haykin (2003).

The final weight matrix after the SOM step is the $m' \times n$ data matrix W'.

$$W' = \begin{bmatrix} w_{11} & \cdots & w_{1m'} \\ \vdots & \ddots & \vdots \\ w_{n1} & \cdots & w_{nm'} \end{bmatrix} \tag{2}$$

2.2. SOM visualization

2.2.1. Unified distance matrix (U-matrix)

The *U*-matrix can be used to visualize the distances between neighboring map units, and thus shows the cluster structure of

Download English Version:

https://daneshyari.com/en/article/6413445

Download Persian Version:

https://daneshyari.com/article/6413445

Daneshyari.com