Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Hydrology 324 (2006) 1–9
Journal of Hydrology
www.elsevier.com/locate/jhydrol

Identification of homogeneous regions for regional frequency analysis using the self-organizing map

Gwo-Fong Lin *, Lu-Hsien Chen

Department of Civil Engineering, National Taiwan University, Taipei 10617, Taiwan, ROC

Received 17 July 2003; revised 4 September 2005; accepted 22 September 2005

Abstract

In this paper, the self-organizing map (SOM) is applied to identify the homogeneous regions for regional frequency analysis. First, the algorithm and structure of the SOM are presented. Then the experimental design is applied to test the cluster accuracy of the SOM, the *K*-means method and Ward's method. These three clustering methods are tested on experimental data sets where the amount of cluster dispersion and the cluster membership are controlled and known. Among the three clustering methods, the results show that the SOM determines the cluster membership more accurately than the *K*-means method and Ward's method. Finally, the SOM is applied to actual rainfall data in Taiwan to identify homogeneous regions for regional frequency analysis. A two-dimensional map indicates that the rain gauges can be grouped into eight clusters. A heterogeneity test indicates that the eight regions are sufficiently homogeneous. Moreover, the results show that the SOM can identify the homogeneous regions more accurately as compared to the other two clustering methods. Because of unsupervised learning, the SOM does not require the knowledge of corresponding output for comparison purposes. In addition, the SOM is more robust than the traditional clustering methods. Therefore, the SOM is recommended as an alternative to the identification of homogeneous regions for regional frequency analysis.

Keywords: Self-organizing map; Homogeneous region; Cluster analysis; Regional frequency analysis

1. Introduction

It is very important to estimate the frequencies and magnitudes of extreme nature events such as floods, rainstorms and droughts. Because the extreme events are rare and the data record is often short, estimation of the frequencies of extreme events is difficult. When data at a given location are insufficient for a reliable estimation of the quantiles, a regional frequency analysis must be performed.

In the process of regional frequency analysis, the sites must be assigned to homogeneous regions, because approximate homogeneity is required to ensure that regional frequency analysis is more accurate than at-site analysis (Hosking and Wallis, 1997; Lin and Chen, 2003). When many sites are involved in a regional frequency analysis, the identification of homogeneous regions is usually

* Corresponding author. Fax: +886 2 2363 1558.
E-mail address: gflin@ntu.edu.tw (G.-F. Lin).

0022-1694/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhydrol.2005.09.009

the most difficult stage and requires the great amount of subjective judgment. Cluster analysis has been successfully used to identify homogeneous regions for regional frequency analysis. Cluster analysis is a standard method of statistical multivariate analysis, and it can reduce large and complex data sets to a small number of data groups where members of a group share similar characteristics.

Regarding the identification of homogeneous regions for use in regional frequency analysis, several relevant studies available in the literature are reviewed in this paper. Mosley (1981) applied cluster analysis to delimitate the hydrological regions in New Zealand, but the regions were not for use in frequency analysis. Burn (1989) used the *K*-means method to derive regions for regional flood frequency analysis. Guttman (1993) and Guttman et al. (1993) used Ward's method and *L*-moments in the determination of regional precipitation climates. Lecce (2000) also used the *K*-means method to examine spatial variations in the timing of flooding in the southeastern United States. Smithers and Schulze (2001) regionalized 172 recording rainfall stations in South Africa using Ward's method. Other examples of the use of cluster analysis in forming hydrological or climatological regions are reported in the literature (Richman and Lamb, 1985; Nathan and McMahon, 1990; Fovell and Fovell, 1993).

Among the aforementioned studies, the *K*-means method and Ward's method are the most frequently used. The *K*-means method is the best known of the nonhierarchical clustering methods. The values of seeds have a great influence on the quality of clustering using a *K*-means clustering or a related technique. When the number of clusters is too large, there is probably no training data in the cluster. In addition, no objective method to determine the number of clusters is another disadvantage. Another clustering method, Ward's method, is an agglomerative hierarchical clustering procedure. Ward's method tends to join clusters that contain a small number of sites, and it is strongly biased when the clusters have roughly the same number of sites. Furthermore, like the *K*-means method, Ward's method does not determine how many clusters actually exist in the data. The *K*-means method and Ward's method can accommodate the large sample sizes. However, the two methods must specify the number of clusters in advance. Unfortunately, the number of clusters is generally unknown in advance. Other clustering methods for determining homogeneous regions also have the same problems. These problems must be solved, and artificial neural networks can present a solution.

Artificial neural networks are currently a popular tool to deal with large amounts of complex data. There are many kinds of artificial neural networks categorized by their learning process or by the network structure. The artificial neural network used in this study is the self-organizing map (SOM) introduced by Kohonen, 1990, 1995). The SOM can project high-dimensional input space on a low-dimensional topology so as to allow the number of clusters to be determined by inspection. It is an example of a competitive network. The SOM was first used as an information-processing tool in the fields of speech and image recognition. More recently, the SOM is widely investigated and reported because of its close ties to biological nervous systems, its simplicity, and the wide variety of problem areas to which it might be applied (Wang et al., 1996; Orwig et al., 1997; Tokutaka et al., 1999; Michaelides et al., 2001; Tennant and Hewitson, 2002; Lin and Chen, 2005). These advantages, coupled with the unsupervised nature of its learning algorithm, have rendered the SOM an attractive alternative for solving various problems that traditionally have been the domain of conventional statistical and operational research techniques. Chen et al. (1995) demonstrated that the SOM is a superior clustering technique and that its relative advantage over conventional techniques increases with higher levels of relative cluster dispersion in the data. Mangiameli et al. (1996) showed that the SOM performed the best when compared to seven other hierarchical clustering methods.

The objective of this paper is to identify homogeneous regions for regional frequency analysis using the SOM. First, the algorithm and architecture of the SOM are presented. Then the SOM is compared with two traditional clustering methods, the *K*-means method and Ward's method, using the controlled experimental data. Finally, the SOM is applied to actual rainfall data in Taiwan to identify homogeneous regions for regional frequency analysis.