Journal of Hydrology 521 (2015) 232–244

Contents lists available at ScienceDirect

Journal of Hydrology

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

Delineation of homogeneous regions for regional frequency analysis using statistical depth function

H. Wazneh [a,*], F. Chebana [a], T.B.M.J. Ouarda [b]

[a] INRS-ETE, 490 rue de la Couronne, Québec (QC), G1K 9A9, Canada
[b] Institute Center for Water and Environment (iWATER), Masdar Institute of Science and Technology, P.O. Box 54224, Abu Dhabi, United Arab Emirates

ARTICLE INFO

Article history:
Received 31 July 2014
Received in revised form 18 November 2014
Accepted 24 November 2014
Available online 10 December 2014
This manuscript was handled by Andras Bardossy, Editor-in-Chief, with the assistance of Ashish Sharma, Associate Editor

Keywords:
Ungauged basins
Homogeneous sub-region
Regional frequency analysis
Statistical depth function
Clustering analysis

SUMMARY

The aim of regional frequency analysis (RFA) is to estimate extreme hydrological events at sites where little or no hydrological data are available. The delineation of sub-regions and the regional estimation within these sub-regions are the two main steps of RFA. As currently practiced, the delineation step is unrobust and subjective. To overcome these limitations, the present paper aims to propose a new robust approach for delineating homogeneous sub-regions. The proposed approach is objective and based on the concept of depth function. A data set from three geographical regions in the North-West of Italy is used to apply and compare the proposed approach with a traditional one. Results indicate that the proposed Depth-based approach leads to more homogeneous sub-regions in terms of *H* heterogeneity measure, and leading to more efficient quantile estimations in terms of relative bias and relative root mean square error, than those obtained by the traditional approach.

1. Introduction

Hydrological data records are mainly short and are not always available in the desired site. Consequently, at-site frequency analysis is not always accurate or even possible at the sites of interest. Regional frequency analysis (RFA) is used to estimate extreme hydrological events at sites where little or no hydrological data are available (e.g., Basu and Srinivas, 2014; Haddad et al., 2014; Ouarda et al., 2000; Reed et al., 1999). Transfer of the available data from gauged sites, within a homogeneous region, to the target (ungauged) site is the main idea behind the RFA. The two main steps of RFA are (i) the identification of homogeneous hydrological sub-regions and (ii) the regional estimation within these sub-regions (e.g., Wazneh et al., 2013b).

The identification of sub-regions has received important consideration in hydrology, but no common methodology has been developed. Thus, various methods of defining sub-regions can be found in the literature, leading to geographically contiguous regions, geographically non-contiguous regions, or hydrological neighbourhoods (Ouarda et al., 2001). Using non-contiguous regions was recommended in the literature (e.g., Haddad and Rahman, 2012; Ouarda et al., 2008). Cluster analysis based on site characteristics is one of the most practical methods used to define the non-contiguous regions (Hosking and Wallis, 1997).

Cluster analysis groups sites based on a distance (measure) reflecting the similarity amongst a set of attributes of the gauging site (e.g., basin area, latitude) (Rao and Srinivas, 2006a). The identified regions highly depend on the choice of the set of attributes (Castellarin et al., 2001; Oudin et al., 2010). Generally, the set of attributes used for RFA approaches includes: (i) physiographic catchment characteristics such as drainage area, average basin slope, main stream slope, stream length (e.g., Acreman and Sinclair, 1986); (ii) geographical location attributes such as latitude, longitude and altitude of catchment centroid (e.g., Burn and Goel, 2000); (iii) measures of basin response time such as basin lag or time-to-peak (e.g., Potter and Faulkner, 1987); (iv) meteorological factors such as storm direction, mean annual rainfall, precipitation intensities (e.g., Chebana and Ouarda, 2008); and (v) at-site flood statistics such as L-moments or other statistical measures calculated from the available flow series (e.g., Wazneh et al., 2013a). A combination of two or more of the above variables may also constitute an attribute in a cluster analysis (e.g., Bargaoui et al., 1998; Nathan and McMahon, 1990).

* Corresponding author. Tel.: +1 (418) 654 2530x4468.
E-mail address: hussein.wazneh@ete.inrs.ca (H. Wazneh).

http://dx.doi.org/10.1016/j.jhydrol.2014.11.068
0022-1694/

Several clustering algorithms are available in the statistical literature (Johnson and Wichern, 2002). Clustering algorithms used for the delineation of sub-regions in RFA can be broadly classified into two categories: hierarchical and partitional clustering (Rao and Srinivas, 2006b). The hierarchical category includes single linkage, complete linkage, average linkage and Ward method (e.g., Baeriswyl and Rebetez, 1997; Bhaskar and O'Connor, 1989; Chiang et al., 2002). The partitional category includes *k*-means and fuzzy *c*-means (e.g., Bargaoui et al., 1998; Rao and Srinivas, 2003). Ward hierarchical method is the most commonly used in RFA. In fact, this method tends to delineate sub-regions approximately equivalent in size, and is thus considered more convenient in the context of regionalizing flood data (Hosking and Wallis, 1997).

In the context of RFA, the application of the above clustering approaches in the delineation step faces two drawbacks. First, these approaches are based on distance measures (e.g., Ward or linkage) and/or use non robust statistics (e.g., *k*-means), and making the delineation results sensitive to noise and to outliers (Ilorme and Griffis, 2013; Jörnsten, 2004). Second, some of these approaches require a preselection of the number of sub-regions (e.g., *k*-means), which makes the delineation step subjective and depends on the user choice.

The aim of the present work is to identify sub-regions for RFA with a particular focus on the formation of sub-regions that can be used for estimating extreme flow quantiles for ungauged sites. More precisely, in this study, a new robust approach for delineation of hydrological sub-regions based on the notion of data depth (Tukey, 1975) is presented and applied. The proposed approach uses, as a starting step, a traditional approach (such as Ward method) to form initial sub-regions. Then, the sites of the initial sub-regions are redistributed in a manner that maximises their depth values (see Section 3). The proposed approach determines objectively the number of homogeneous sub-regions using a preselected criterion such as the *H* heterogeneity measure (Hosking and Wallis, 1997).

Wazneh et al., 2013a introduced statistical depth function in regional estimation through the index-flood model. To delineate the homogeneous sub-regions, they used traditional Ward method. They showed that employing depth function in the estimation step provides improved results. However, as argued by the authors, it is more appropriate to consider depth function in the delineation step as well as in the estimation which ensures compatibility between the two main RFA steps (delineation and estimation) and could lead to better results. Therefore, it is necessary to develop a Depth-based regional delineation.

Several depth functions are available in the literature and have been used for generalizing many univariate statistical methods to the multivariate set-up (e.g., Chen et al., 2009; Donoho and Gasko, 1992). In particular, Jörnsten (2004) introduced the notion of depth functions in the clustering approach. In this study, the author defined the maximum depth clustering, where an observation is assigned to the cluster with which it has the maximum depth value. From a simulation study, the author shows that the depth clustering approach is robust to noise and outliers. She shows that employing depth functions improves clustering accuracy. After Jörnsten (2004), statistical depth functions are employed in a number of clustering and classification approaches (e.g., Dutta and Ghosh, 2012; Ghosh and Chaudhuri, 2005).

This paper is organised as follows. Section 2 assembles the various elements of the background needed to introduce the proposed approach and compare it with the traditional Ward approach. The proposed approach in its general form is described in Section 3. Its application in a real world data set is presented in Section 4. The last section is devoted to the conclusions of this work.

2. Background

This section briefly presents the background material required to introduce and apply the delineation of sub-regions using depth functions. In addition, this section includes several basic concepts as well as a summary of the main steps of the index-flood model used as a regional estimation model in order to quantify the performance of the proposed approach.

2.1. Ward method

In this section, the Ward method, commonly used to delineate homogeneous sub-regions in RFA analysis, is presented. This method is considered in this paper for comparison purposes.

Ward method (Ward, 1963) is a hierarchical algorithm which initially begins with each site serving as its own sub-region. The algorithm successively merges sub-regions using an analysis of variance approach in which the similarity amongst sites in a sub-region is measured in terms of the Error Sum of Squares (*ESS*). More formally, for a sub-region *r* containing *s* sites, where the flood regime is represented by *p* attributes $X = (X_1, X_2, \ldots, X_p)$, the *ESS* is given by:

$$ESS_r = \sum_{j=1}^{s} (X_j - \overline{X}_r)' (X_j - \overline{X}_r) \quad (1)$$

where $X_j = (X_{j1}, X_{j2}, \ldots, X_{jp})$ is a vector of the attributes at site *j*, and $\overline{X}_r$ is a vector of the means of the attributes within the sub-region *r*. At each step, ESS_r is computed for the hypothetical merger of any two sub-regions, and the actual mergers chosen to occur are those which minimise the increase in the total *ESS* across all sub-regions. A dendrogram is commonly used to illustrate the mergers made at successive levels, where the vertical axis represents the value of the *ESS*.

2.2. Spatial depth function

Data depth is a quantitative measure of how central (or deep) a point is with respect to a data set or a distribution in a multivariate framework. This gives us a central outward ordering of multivariate data points and gives rise to new ways to quantify the many complex multivariate features of the underlying multivariate distribution (Li et al., 2012; Liu et al., 1999). The depth functions were first introduced by Tukey (1975). They are employed in many research fields including water sciences (e.g., Bárdossy and Singh, 2008, 2011; Chebana and Ouarda, 2008; Krauße and Cullmann, 2012; Krauße et al., 2012). For a given cumulative distribution function *F* on $\Re^d (d \geqslant 1)$, a depth function is any non-negative bounded function which has a number of convenient properties. These properties fit well the RFA requirements and constraints (Chebana and Ouarda, 2011).

Several types of depth functions have been developed e.g., half space, projection, simplicial, spatial and Mahalanobis depth functions. Several depth functions, practical in two dimensions, become impractical where the dimensionality increases e.g., simplicial depth and half space (Hugg et al., 2006). The spatial depth function was used in this study because of its convenient properties. First, spatial depth is non-zero outside the convex hull of the data set and therefore can be used to clustering sites of the region (see Section 3.1). Second, it is fast and easy to compute in any dimension, contributing to its application for the study of large high-dimensional data sets, such clustering studies.

The spatial depth of point *x* is the amount of probability mass needed at *x* to make it the multivariate median (spatial median) of the data. Formally, the spatial depth of point *x* on $\Re^d (d \geqslant 1)$ is: