Journal of Hydrology 524 (2015) 30–43

Contents lists available at ScienceDirect

Journal of Hydrology

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

Analytical approach to quantile estimation in regional frequency analysis based on fuzzy framework

Bidroha Basu [1], V.V. Srinivas *

Department of Civil Engineering, Indian Institute of Science, Bangalore 560 012, India

ARTICLE INFO

Article history:
Received 20 September 2014
Received in revised form 31 December 2014
Accepted 9 February 2015
Available online 17 February 2015
This manuscript was handled by Andras Bardossy, Editor-in-Chief, with the assistance of Attilio Castellarin, Associate Editor

Keywords:
Regionalization
Regional flood frequency analysis
Fuzzy cluster analysis
Mid-Atlantic water resources region, USA

SUMMARY

Regional frequency analysis is widely used for estimating quantiles of hydrological extreme events at sparsely gauged/ungauged target sites in river basins. It involves identification of a region (group of watersheds) resembling watershed of the target site, and use of information pooled from the region to estimate quantile for the target site. In the analysis, watershed of the target site is assumed to completely resemble watersheds in the identified region in terms of mechanism underlying generation of extreme event. In reality, it is rare to find watersheds that completely resemble each other. Fuzzy clustering approach can account for partial resemblance of watersheds and yield region(s) for the target site. Formation of regions and quantile estimation requires discerning information from fuzzy-membership matrix obtained based on the approach. Practitioners often defuzzify the matrix to form disjoint clusters (regions) and use them as the basis for quantile estimation. The defuzzification approach (DFA) results in loss of information discerned on partial resemblance of watersheds. The lost information cannot be utilized in quantile estimation, owing to which the estimates could have significant error. To avert the loss of information, a threshold strategy (TS) was considered in some prior studies. In this study, it is analytically shown that the strategy results in under-prediction of quantiles. To address this, a mathematical approach is proposed in this study and its effectiveness in estimating flood quantiles relative to DFA and TS is demonstrated through Monte-Carlo simulation experiments and case study on Mid-Atlantic water resources region, USA.

1. Introduction

Estimation of design quantiles of floods at sparsely gauged/ungauged target sites in river basins is often necessary for various applications in water resources engineering. This is accomplished using regional frequency analysis (RFA) that involves identification of a region (group of gauged watersheds) resembling watershed of the target site and use of information pooled from the region as the basis to arrive at the quantile estimate(s) for the target site. Various approaches to RFA have been developed over the past five decades that differ in strategies for region identification and pooling regional information. Relevant literature can be found in Institute of Hydrology (1999), Rao and Srinivas (2008), Blöschl et al. (2013) and Hrachowitz et al. (2013).

Conventional regionalization approaches attempt to delineate watersheds available in the study area into disjoint regions (-groups) such that watersheds in a region completely resemble in terms of attributes that depict mechanism underlying generation of floods. Delineation of such regions is not possible in real world scenario, as it is rare to find watersheds that completely resemble each other. Since the end of twentieth century, fuzzy clustering procedures gained recognition for use in regionalization of watersheds for RFA (e.g., Bargaoui et al., 1998; Hall and Minns, 1999; Jingyi and Hall, 2004; Rao and Srinivas, 2006, 2008; Srinivas et al., 2008; Basu and Srinivas, 2014), as those procedures allow watersheds to partially resemble each other. Formation of regions and quantile estimation requires discerning information from fuzzy-membership matrix obtained based on the procedures. But, currently there are no effective procedures for quantile estimation by utilizing information available in fuzzy-membership matrix that corresponds to scenario where sites do not fully resemble one another. Often modelers defuzzify (harden) membership matrix to form disjoint clusters (regions) (Hall and Minns, 1999; Jingyi and Hall, 2004; Güler and Thyne, 2004; Sadri and Burn,

* Corresponding author.
E-mail addresses: bidroha@civil.iisc.ernet.in (B. Basu), vvs@civil.iisc.ernet.in (V.V. Srinivas).
[1] Contributing author.

http://dx.doi.org/10.1016/j.jhydrol.2015.02.026
0022-1694/

2011) and use them as the basis for quantile estimation, which results in loss of information discerned on partial resemblance of watersheds. The lost information cannot be utilized in quantile estimation, owing to which the quantile estimates could have significant error. To avert the loss of information, a threshold strategy was considered in a few studies (Rao and Srinivas, 2006, 2008; Basu and Srinivas, 2014) that aims to (i) form fuzzy clusters based on a threshold specified for fuzzy-memberships, thus averting hardening of membership matrix, and (ii) utilize fuzzy-memberships in quantile estimation. In this study, it is shown through analytical formulations that the use of the threshold strategy results in under-prediction of quantiles. To address this problem, a new mathematical procedure is proposed. Its effectiveness is demonstrated through Monte-Carlo simulation experiments and by application to watersheds in a water resources region of United States.

The remainder of this paper is structured as follows: background of the problem being addressed in this paper is provided in Section 2. Following that, methodology of the proposed mathematical procedure is presented in Section 3. Subsequently, effectiveness of the proposed procedure over defuzzification approach and threshold strategy is demonstrated through Monte Carlo simulation experiments and by application to real world data in Section 4. Finally, summary and concluding remarks are given in Section 5.

2. Background

2.1. Conventional fuzzy clustering approach to regionalization

Let there be N sites (stream gauges) in the study area, each representing a watershed. Application of a fuzzy clustering procedure for regionalization of the study area yields c fuzzy clusters, their centroids and fuzzy-membership matrix containing information on 'degree of belongingness' of each of the watersheds to each of the clusters. Let u_{si} denote fuzzy-membership of site s (=1,…,N) in cluster i (=1,…,c). Traditionally, modelers defuzzify (harden) the clusters using Eq. (1) to form regions, and then pool information in each of the regions for regional frequency analysis. This approach is henceforth referred to as defuzzification approach (DFA).

$$u_{sj} = \max_{1\leqslant i\leqslant c}\{u_{si}\} = 1; \quad u_{si} = 0 \quad \forall i \neq j \tag{1}$$

Eq. (1) implies that each watershed is assigned to one of the clusters (regions) to which it has maximum resemblance. The DFA results in loss of information, as memberships u_{si} for $i \neq j$ are forced to zero, even if they are significant. In literature, there are no prior attempts to quantify the effect of defuzzification on quantile estimates, as exact formulations to estimate true quantiles in fuzzy framework are unknown.

2.2. Conventional procedure to estimate quantile for ungauged site

In order to estimate flood quantile for the target ungauged/sparsely gauged site s', modelers conventionally opt for index-flood approach (Dalrymple, 1960) that involves the following steps.

(1) Assignment of the site s' to one of the c clusters in which it has maximum membership.
(2) Derivation of normalized records for each of the sites in the cluster by dividing peak flows at the site by their mean value (index-flood value).
(3) Identification of an appropriate regional frequency distribution (RFD) to fit the normalized records.
(4) Computation of weighted average of summary statistics (mean, variance and skewness) of normalized records corresponding to all the sites in the cluster, and use of those statistics to estimate parameters of RFD and construct a dimensionless cumulative distribution function $\hat{q}(\cdot)$, which is referred to as growth curve.
(5) Estimation of quantile for the target site s' using index-flood method (Dalrymple, 1960) as,

$$\widehat{Q}^R_{s'}(T) = \overline{Q}_{s'}\hat{q}(T) \tag{2}$$

where $\widehat{Q}^R_{s'}(T)$ is quantile corresponding to T-year return period; $\overline{Q}_{s'}$ is the index-flood value for the target site, which can be estimated by inputting attributes of the target site in multiple linear regression relationship developed between attributes and index-flood value corresponding to all the sites in the cluster.

2.3. Regionalization and quantile estimation using threshold strategy

To avert defuzzification of fuzzy-membership matrix and consequent loss of information, a threshold strategy (TS) was proposed by Rao and Srinivas (2006). The strategy aims to form fuzzy clusters (regions) such that each cluster comprises sites whose memberships in it are greater than or equal to a threshold value $1/c_o$, where c_o denotes optimal number of clusters. In order to estimate flood quantile for the target ungauged/sparsely gauged site, fuzzy-membership $u^o_{s'i}$ of the site in each of the c_o clusters is determined. Following this, the clusters $\{1,\ldots,c^{s'}_o\}$ in which the site has membership greater than or equal to the threshold $1/c_o$ are identified, and growth curves $\{\hat{q}_j(\cdot), j \in \{1,\ldots,c^{s'}_o\}\}$ for those identified clusters are constructed following the steps (2) to (4) given in Section 2.2. Then $\widehat{Q}^R_{s'}(T)$ is estimated using the following equation, instead of Eq. (2).

$$\widehat{Q}^R_{s'}(T) = \overline{Q}_{s'} \sum_{\forall j, j\in\{1,\ldots,c^{s'}_o\}} w_j\hat{q}_j(T);$$
$$\text{where} \quad \overline{Q}_{s'} = \sum_{\forall j, j\in\{1,\ldots,c^{s'}_o\}} w_j\overline{Q}^{(j)}_{s'}$$
$$\text{and} \quad w_j = \left(\frac{u^o_{s'j}}{\sum_{\forall l, l\in\{1,\ldots,c^{s'}_o\}} u^o_{s'l}}\right) \tag{3}$$

where w_j denotes weight assigned to $\hat{q}_j(\cdot)$ in proportion to membership of the target site in jth of the $c^{s'}_o$ clusters to which it belongs. $\overline{Q}_{s'}$ is the index-flood value for the target site s'; $\overline{Q}^{(j)}_{s'}$ is mean peak flow estimate for site s' that is obtained (using information of region j) by inputting attributes of the site s' in multiple linear regression relationship developed between attributes and mean peak flow value corresponding to all the sites in the region j. Analytically it can be shown that use of Eq. (3) results in under-prediction of quantiles.

In fuzzy framework, peak flow $Q_{s'}$ at the target site s' must be considered to be the outcome of a mixture of $c^{s'}_o$ independent random variables $Y_1,\ldots,Y_{c^{s'}_o}$, each of which corresponds to a different cluster (region) in which the target site has membership.

$$Q_{s'} = \overline{Q}_{s'} \sum_{\forall j, j\in\{1,\ldots,c^{s'}_o\}} w_jY_j \tag{4}$$

Growth curve $\hat{q}_j(\cdot)$ denotes cumulative distribution function (CDF) corresponding to a random variable Y_j. Probability of $Q_{s'}$ not exceeding $\widehat{Q}^R_{s'}(T)$ that is denoted by $P\left(Q_{s'} \leqslant \widehat{Q}^R_{s'}(T)\right)$, can be expanded as follows.