Journal of Hydrology 438–439 (2012) 26–38

Contents lists available at SciVerse ScienceDirect

Journal of Hydrology

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

Advances in variable selection methods II: Effect of variable selection method on classification of hydrologically similar watersheds in three Mid-Atlantic ecoregions

H. Ssegane [a], E.W. Tollner [a,*], Y.M. Mohamoud [b], T.C. Rasmussen [c], J.F. Dowd [d]

[a] Department of Biological and Agricultural Engineering, University of Georgia, Athens, GA 30602, USA
[b] Ecosystems Research Division, US Environmental Protection Agency, Athens, GA 30605, USA
[c] Warnell School of Forestry and Natural Resources, University of Georgia, Athens, GA 30602, USA
[d] Department of Geology, University of Georgia, Athens, GA 30602, USA

ARTICLE INFO

Article history:
Available online 2 February 2012
This manuscript was handled by Andras Bardossy, Editor-in-Chief, with the assistance of Attilio Castellarin, Associate Editor

Keywords:
Watershed classification
Hydrological simillarity
Streamflow indices
Causal variable selection
Stepwise regression
Principal component analysis

ABSTRACT

Hydrological flow predictions in ungauged and sparsely gauged watersheds use regionalization or classification of hydrologically similar watersheds to develop empirical relationships between hydrologic, climatic, and watershed variables. The watershed classifications may be based on geographic proximity, regional frameworks such as ecoregions or classification using cluster analysis of watershed descriptors. General approaches used in classifying hydrologically similar watersheds use climatic and watershed variables or statistics of streamflow data. Use of climatic and watershed descriptors requires variable selection to minimize redundancy from a large pool of potential variables. This study compares classification performance of four variable groups to identify homogeneous watersheds in three Mid-Atlantic ecoregions (USA): Appalachian Plateau, Piedmont, and Ridge and Valley. The variable groups included: (1) variables that define watershed geographic proximity; (2) variables that define watershed hypsometry; (3) variables selected using causal selection algorithms; and (4) variables selected using principal component analysis (PCA) and stepwise regression. The classification results were compared to reference watersheds classified as homogeneous using three streamflow indices: Slope of flow duration curve; Baseflow index; and Streamflow elasticity using a similarity index (*SI*). Classification performance was highest using variables selected by causal algorithms (e.g., HITON-MB method, $SI = 0.71$ for Appalachian Plateau, $SI = 0.90$ for Piedmont, and $SI = 0.72$ for Ridge and Valley) compared to variables selected by stepwise regression ($SI = 0.72$ for Appalachian Plateau, $SI = 0.87$ for Piedmont, and $SI = 0.64$ for Ridge and Valley) and PCA ($SI = 0.71$ for Appalachian Plateau, $SI = 0.76$ for Piedmont, and $SI = 0.57$ for Ridge and Valley).

1. Introduction

Development of regional frameworks such as hydrological landscape regions (Wolock et al., 2004) and ecoregions (Omernik and Bailey, 1997) has led to regionalization (Hall and Minns, 1999) of streamflow indices such that observed streamflow at gauged sites can be extrapolated to predict streamflow at ungauged sites in the same physiographic region. The concept of regionalization assumes that watersheds in the same physiographic region have similar hydrological signatures over a long period of time. Regionalization methods include: (1) statistical regionalization, where multiple regression is used to link hydrological responses to physical and climatic attributes (Kokkonen et al., 2003); (2) use of geospatial similarity (Merz and Blöschl, 2004); and (3) use of regional hydrological model parameters (Bastola et al., 2008). Irrespective of the approach used, observed data at gauged sites is used to model underlying hydrological processes at ungauged sites. Although previous studies have shown that geospatial similarity or geographical proximity does not always translate into hydrological similarity (Kokkonen et al., 2003; Acreman and Sinclair, 1986), geographic proximity may infer similarity in climatic conditions and watershed form. Commonly used approaches include those that infer similarity using climatic and watershed variables and those that use streamflow statistics or both.

Chiang et al. (2002) used cluster analysis and 16 streamflow statistics to generate six homogeneous regions from 94 watersheds in Alabama, Georgia, and Mississippi (USA). Kahya et al. (2008) used hierarchical clustering and streamflow patterns to classify 80 watersheds in Turkey. Acreman and Sinclair (1986) used 11

[*] Corresponding author. Tel.: +1 706 202 8193.
E-mail addresses: seganeh@uga.edu (H. Ssegane), btollner@engr.uga.edu (E.W. Tollner), Mohamoud.Yusuf@epamail.epa.gov (Y.M. Mohamoud), trasmuss@uga.edu (T.C. Rasmussen), jdowd@uga.edu (J.F. Dowd).

0022-1694/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhydrol.2012.01.035

watershed variables to classify 168 watersheds in Scotland into 5 homogeneous regions. And, Di Prinzio et al. (2011) used six streamflow statistics to establish reference homogeneous regions and compared results to four alternative classification methods using 12 watershed variables. The challenge with the above approaches is that there are no universally accepted similarity metrics (Wagener et al., 2007). Also, the watershed classification results depend on watershed descriptors used or the effectiveness of the variable selection methods.

On the choice of streamflow indices, Sawicz et al. (2011) suggest six streamflow metrics that define the different hydrologic functions of watersheds as possible universal metrics. The metrics include runoff ratio, flow duration curves, baseflow index, streamflow elasticity, ratio of snow days, and rising limb density. However, streamflow indices cannot be used to determine hydrological similarity of ungauged watersheds. On the choice of watershed descriptors, the most used variable selection methods are principal component analysis (PCA) (Salas et al., 2010; Ma et al., 2010; Alcázar and Palau, 2010) and stepwise regression analysis (SRA) (Barnett et al., 2010; Gong et al., 2010; Peña-Arancibia et al., 2010). The conceptual basis of both approaches is not causality between response and explanatory variables. Stepwise regression analysis focuses on minimization of the predictive error while principal component analysis focuses on dimensional reduction (data extraction) by projecting high dimension data onto a low dimension space while maintaining the most relevant information.

Causal relationships between response and explanatory variables can be discovered by Bayesian networks. Bayesian networks consist of directed acyclic graphs whose nodes represent random variables and the edges conditional probabilities (Jensen and Nielsen, 2007; Karimi and Hamilton, 2009; Meganck et al., 2006). Therefore, the implied causation by this approach is probabilistic causation based on the theory that causes increase or change the probabilities of their effects such that the conditional probability of an effect given its cause is greater than the probability of the effect in absence of the cause (Hitchcock, 2010; Suppes, 1970; Cartwright, 1979). Thus, the possibility of event *A* occurring given that event *B* occurred is higher if event B causes event A and vice-versa. Some of the algorithms that implement causal variable selection include: Grow-Shrink, GS (Margaritis and Thrun, 1999); interleaved Incremental Association Markov Boundary with PC algorithm, interIAMBnPC (Tsamardinos et al., 2003); Local Causal Discovery, LCD2 (Cooper, 1997); and HITON Markov Blanket, HITON-MB (Aliferis et al., 2003). For a brief description of the methods, the readers should refer to the first part of this study (reference for part I).

Therefore, the objective of the second part of the study is to compare the effectiveness of determining hydrologically similar watersheds using variables selected by causal algorithms (GS, interIAMBnPC, LCD2, and HITON-MB), stepwise regression analysis, principal component analysis, variables of geographical proximity, and watershed hypsometry in three Mid-Atlantic ecoregions: Appalachian Plateau, Piedmont, and Ridge and Valley (USA). The variable groups selected for comparison included: (1) variables that define watershed geographical proximity; (2) variables that define watershed hypsometry; (3) variables selected using causal selection algorithms; and (4) variables selected using principal component analysis (PCA) and stepwise regression. Hence, the focus of this study is on the effect of different variable selection methods on watershed classification while many previous studies have focused on different clustering or regionalization methods using the same set of variables.

We hypothesize that although hydrological similarity between watersheds in the same ecoregion is high when compared to watersheds from different ecoregions, all watersheds in the same ecoregion may not hydrologically behave in a similar manner. Therefore, the study used three streamflow indices: (1) slope of a flow duration curve (FDC); (2) the baseflow index (BFI); and (3) streamflow elasticity (SFE) with *k*-means clustering to classify reference homogeneous watersheds for each ecoregion. Watersheds classified using streamflow indices were considered to be the true hydrologically similar watersheds (reference watersheds) for each ecoregion. Then the ability of the four watershed variable groups to generate the exact homogeneous watersheds for the Appalachian Plateau, Piedmont, and Ridge and Valley were examined using a similarity index. The *a priori* assumption is that watershed classification using variables that typify the cause and effect relationship with the streamflow indices should give highest similarity when compared to reference watersheds.

The relevance of this approach was to emphasize the dependence and accuracy of watershed classification results on the variables used for classification. The interest in geographical proximity of watersheds is because proximity may infer similar climatic conditions and watershed form. While the interest in watershed hypsometry is based on the role of topography in hydrological processes. Stieglitz et al. (1997) highlighted the role of topography on soil moisture distribution, timing of discharge, and partitioning of streamflow into direct runoff and baseflow. Also, Vivoni et al. (2008) showed that total runoff reduced as the watershed hypsometric form changed from convex to concave. Therefore, this study also evaluates whether statistics of a hypsometric curve are adequate representatives of topography to differentiate hydrologic behavior across the three ecoregions.

2. Methods

2.1. Study area and data

Data used in this study covers three Mid-Atlantic physiographic regions (ecoregions) within USA (Fig. 1); the Appalachian Plateau (26 watersheds), the Piedmont (25 watersheds), and the Ridge and Valley (29 watersheds). Streamflow data used spanned the same 42 years of 1966–2007 epoch across all watersheds. Fig. 2 depicts topographic differences of headwaters of representative watersheds from each ecoregion. The watersheds were selected from Hydro-Climatic Data Network (HCDN) dataset (Slack and Landwehr, 1992) with emphasis on low extent of urbanization and minimum surface storage. For detailed description of the climatic and watershed descriptors used in this second part of the study, the reader is referred to part I of the study or Table 2.

2.2. Streamflow metrics

The common measures of watershed homogeneity or hydrological similarity analysis involve use of streamflow statistics (Kahya et al., 2008; Srinivas et al., 2008; Castellarin et al., 2008; Patil and Stieglitz, 2011). Three measures of watershed function signature were used to define hydrological similarity for watersheds in the same ecoregion. The measures included the slope of a flow duration curve (FDC), the baseflow index (BFI), and the streamflow elasticity. These three indices are a subset of six indices recommended by Sawicz et al. (2011). The choice of the three streamflow metrics was based on: (1) adequate representation of the watershed hydrologic response by the three metrics (refer to Sections 2.2.1, 2.2.2, 2.2.3); (2) use of fewer variables minimizes challenges of using high dimension data for unsupervised learning such as clustering (Fern and Brodley, 2003; Ding et al., 2002; Müller et al., 2009); and (3) the three metrics were easily extracted from readily available data compared to extracting all six indices. Watersheds classified as hydrologically homogeneous based on