Journal of Hydrology 438–439 (2012) 16–25

Contents lists available at SciVerse ScienceDirect

Journal of Hydrology

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

Advances in variable selection methods I: Causal selection methods versus stepwise regression and principal component analysis on data of known and unknown functional relationships

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, Driftmier Engineering Center, Athens, GA 30602, USA
[b] Ecosystems Research Division, US Environmental Protection Agency, 960 College Station Road, 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, Geography–Geology Building, Athens, GA 30602, USA

ARTICLE INFO

Article history:
Available online 20 January 2012
This manuscript was handled by Andras Bardossy, Editor-in-Chief, with the assistance of Attilio Castellarin, Associate Editor

Keywords:
Causal variable selection
Stepwise regression
Principal component analysis
Watershed variables
Streamflow percentiles

SUMMARY

Hydrological predictions at a watershed scale are commonly based on extrapolation and upscaling of hydrological behavior at plot and hillslope scales. Yet, dominant hydrological drivers at a hillslope may not be as dominant at the watershed scale because of the heterogeneity of watershed characteristics. With the availability of quantifiable watershed data (watershed descriptors and streamflow indices), variable selection can provide insight into the dominant watershed descriptors that drive different streamflow regimes. Stepwise regression and principal components analysis have long been used to select descriptive variables for relating runoff to climate and watershed descriptors. Questions have remained regarding the robustness of the selected descriptors. This paper evaluates five new approaches: Grow-Shrink, GS; a variant of Incremental Association Markov Boundary, interIAMBnPC; Local Causal Discovery, LCD2; HITON Markov Blanket, HITON-MB; and First-Order Utility, FOU. We demonstrate their performance by quantifying their accuracy, consistency and predictive potential compared to stepwise regression and principal component analysis on two known functional relationships. The results show that the variables selected by HITON-MB and the first-order utility are the most accurate while variables selected by Stepwise regression, although not accurate have a high predictive potential. Therefore, a model with high predictive power may not necessary represent the underlying hydrological processes of a watershed system.

1. Introduction

Hydrological predictions in ungauged watersheds include estimation of hydrological responses in watersheds that have no flow measuring instruments, watersheds with fewer gauges compared to watershed size, and watersheds that are gauged but have few years of data (Sivapalan et al., 2003). These predictions are based on climatic inputs, land use and land cover, soil and physical descriptors, and watershed topography. Hydrological predictions are relevant to analysis of changes due to deforestation, urbanization, stream withdrawals, and installation and operation of reservoirs. Several methods are used to model hydrological behavior in ungauged watersheds. The methods include statistical regionalization (Kokkonen et al., 2003) and the use of regional hydrological model parameters (Bastola et al., 2008). Both approaches use observed data at gauged sites to conceptualize and derive underlying hydrological processes for predictions at ungauged sites.

Examination of 42 published papers (e.g. Alcazar et al., 2008; Johnston and Shmagin, 2008; Mohamoud, 2008; Sando et al., 2009) identified 72 unique topographic variables, 66 climatic variables, 98 soil variables, and 15 land use and land cover variables used by different researchers. The selection of relevant variables and whether their effect is positive or negative vary: (1) from region to region; (2) depending on the initial watershed variables used; (3) depending on the conceptualization by different researchers of what constitutes relevant variables; and (4) depending on the variable selection method. Fig. 1 summarizes some of the watershed descriptors used in the above studies. These studies, though not extensive, provide a basis for *a priori* assumptions on the role of topography, climate, land use, and soil descriptors at different flows.

Although the majority of the variables are statistically redundant, the challenge is to devise approaches that minimize variable redundancy and identify relevant variables that characterize the full behavior of flow regimes on a regional basis. Commonly used

* Corresponding author. Tel.: +1 706 542 3047.
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.2012.01.008

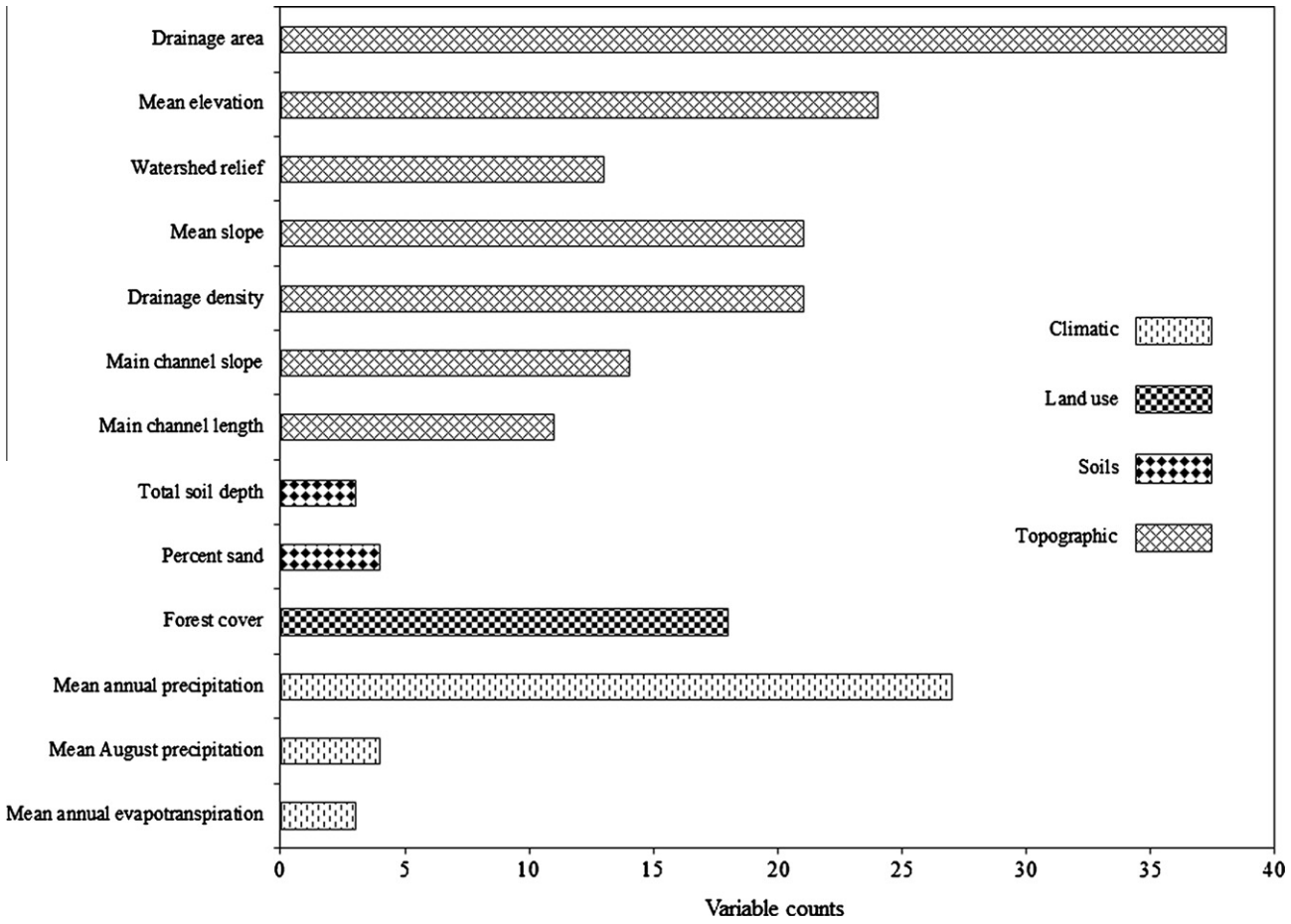

Fig. 1. Frequency counts of most selected variables, based on literature review of 42 studies between 1989 and 2009. Shading indicates category of watershed variable. The four variable categories used in this study include; (1) Topography, (2) Soils, (3) Landuse and Landcover, and (4) Climate.

approaches include stepwise regression (Heuvelmans et al., 2006; Brandes et al., 2005; Barnett et al., 2010; Gong et al., 2010; Peña-Arancibia et al., 2010) and principal component analysis (Salas et al., 2010; Ma et al., 2010; Morris et al., 2009; Alcazar and Palau, 2010). Stepwise regression seeks to minimize the prediction error while principal component analysis focuses on dimension reduction which may not utilize information from the response variable. Both approaches perform well (high coefficient of determination; $R^2 \geqslant 0.8$) but are susceptible to the elimination of relevant variables. Neither method is structured to derive causal associations between dependent (response) and independent (explanatory) variables. Also, use of a limited pool of independent variables may result in selection of irrelevant variables as relevant in absence of other relevant variables; a concept referred to as Simpson's paradox (Whittaker, 1990). This paradox states that two variables may be marginally independent in absence of a third variable but become dependent when conditioned on the third variable.

Advancements in the fields of artificial intelligence, machine learning, and data mining, in addition to increased computational speed have led to development of methods that seek to infer causal associations between explanatory and response variables. Causal relationships between response and explanatory variables can be discovered using 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 found using Bayesian networks is a probabilistic causation based on the precept 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).

Some of the causal methods include: Grow-Shrink, GS (Margaritis and Thrun, 1999); a variant of Incremental Association Markov Boundary (IAMB), interIAMBnPC (Tsamardinos et al., 2003); Local Causal Discovery, LCD2 (Cooper, 1997); HITON Markov Blanket, HITON-MB (Aliferis et al., 2003a); and First Order Utility, FOU (Brown, 2009). The first four methods seek to select causal variables by reconstructing a Markov blanket of the response variable based on probabilistic definition of causation and variable relevance, while the fifth method (FOU) uses mutual information to derive variable relevance, redundancy, and conditional redundancy. The four causal selection algorithms have two major phases; the growing phase, where variables are added to a Markov blanket (MB) and a shrinking phase, where false positives are removed. The GS statically orders the variables based on their association with the response variable given the empty MB and then admits into MB the variable in the ordering that is not conditionally independent with response given the current MB. The IAMB is similar to GS; however, each time a new variable enters a candidate MB, the algorithm reorders the variables based on the updated conditional independence test. The interIAMBnPC interleaves the growing phase of IAMB with the shrinking phase; however it replaces the shrinking phase of IAMB with the Peter-Clark algorithm (Spirtes et al., 2000). The LCD2 implements five tests of dependence and one test of independence between an instrumental variable, the response variable, and the variable of