Input variable selection for hydrological predictions in ungauged catchments: with or without clustering?

Nilay Dogulu1, Inci Batmaz2, and Elcin Kendel1

1 Middle East Technical University, Department of Civil Engineering, Water Resources Laboratory, Ankara, Turkey
2 Middle East Technical University, Department of Statistics, Ankara, Turkey

IVS methods

Mean Partial Correlation (PCIS)

CAMELS

Where?

CATCHMENT Attributes and Microbiology for Large-sample Studies Ador et al. 2017, HESS

671 watersheds across continental USA

(Parameter / less impacted by anthropogenic changes)

Large-scale hydrology can help to gain useful insights on how these significant predictors change over different runoff signatures representing particular hydrological conditions as well as within different groups of similar catchments.

Research Objective

To explore the added value of clustering for input variable selection

Methodology

Clusterings of catchments using available topography, soil, geology, vegetation and climate attributes

Input Variable Selection (IVS) for each cluster & hydrological attribute

Partial Mutual Information (PMI)

Partial Correlation Input Selection (PCIS)

Iterative Input Selection (IS)

Data

Clustering is performed on input space consisting of 31 (numeric) variables representing catchment characteristics

IVS is important for hydrological predictions in ungauged catchments! And clustering of input data space yields value!

• Climate and topography related variables are more frequently selected compared to soil, geology and vegetation attributes. However, there are cases where geology becomes more important as a predictor of the system.
• We found that selected variables for predicting hydrological attributes are not the same for different clusters.
• Moreover, selected variables for a given cluster vary depending on the hydrological attributes of interest.
• More in-depth analysis of IVS results is in progress.
• The effect of clustering method choice could be further explored.
• Next step: using data-driven models for each cluster to make hydrological predictions on validation set.

Contact

Elcin Kendel

Email: incibatmaz@metu.edu.tr

Inci Batmaz

Email: incibatmaz@metu.edu.tr

Nilay Dogulu

Email: nilaydogulu@metu.edu.tr

Summary

Partial Correlation Input Selection (PCIS)

• The partial correlation input selection algorithm is based on partial correlation coefficient.
• A PCIS algorithm is used to estimate the strength of the relationship between inputs and output and to produce a multiple linear regression model’s list of variables. A PCIS algorithm is used to estimate the selection of the most statistically significant variables.
• The list of selected variables and the selection criteria are available through the CAMELS project (Ador et al., 2017).

Iterative Input Selection (IS)

• A stepwise forward regression method is used to select input variables from the CAMELS data set.
• The IS algorithm is used to estimate the selection of the most statistically significant variables.
• The list of selected variables and the selection criteria are available through the CAMELS project (Ador et al., 2017).

Catches in north and mid-western coastal regions are identified by:

*Clustering 1 and 8, respectively.*

Lesser representing emergent cluster with WET climate (i.e., aridity index ≤5) can be seen in the Budipo Cume.

By looking at the Budipo Curve, one can say that the differences are annual water balance changes.

K-means clustering

• Most well-known traditional clustering method based on a center-based approach (e.g., K-means) is applied here to form ungauged catchments into clusters of similar hydrological behavior.
• Several different clustering algorithms are used, and the best performing one is selected based on the initial cluster of the dataset.

Input Variable Selection (IVS)

• Information theory based filter algorithm developed by Sharma (2000), later modified by Inci Batmaz et al. (2012). The algorithm is used to estimate the information gained from the data.
• Extra-Trees (Geurts et al., 2006) is used in the last step to select the final predictor. The Extra-Trees is used after removing the effects of other variables. Pearson correlation coefficient is used to estimate the additional dependence the new predictor can add to the existing prediction model.


© 2019 Inci Batmaz, Elcin Kendel, and Nilay Dogulu. This is an open access article distributed under the Creative Commons Attribution 4.0 International License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
Input variable selection for hydrological predictions in ungauged catchments: with or without clustering?

Nilay Dogulu (1), Inci Batmaz (2), and Elcin Kentel (3)
(1) Middle East Technical University (METU), Civil Engineering Dept., Water Resources Laboratory, Ankara, Turkey (ndogulu@metu.edu.tr), (2) Middle East Technical University (METU), Statistics Dept., Ankara, Turkey (ibatmaz@metu.edu.tr), (3) Middle East Technical University (METU), Civil Engineering Dept., Water Resources Laboratory, Ankara, Turkey (ekentel@metu.edu.tr)

A key step in data-driven environmental modelling, including for hydrological purposes, is input variable selection (IVS) to ensure that the least number of variables with minimum redundancy are used to characterize the inherent relationship between inputs and outputs. Hydrological predictions in ungauged catchments is one such area where the information on influential predictors of runoff signatures guides in understanding dominant controls of meaningful information transfer from gauged to ungauged locations (i.e. regionalization). This understanding is valuable especially for the analysis of hydrological similarity among ungauged catchments, e.g. to identify reference (donor) catchment(s). Large-sample hydrology can help to gain useful insights on how these significant predictors change over different runoff signatures representing particular hydrological conditions as well as within different groups of similar catchments. This study explores the added value of clustering for input variable selection in the case of catchments across continental USA using the CAMELS dataset (Addor et al., 2017). We employ the method of k-means clustering on the input space consisting of topography, soil, geology, vegetation and climate attributes as a way to deal with heterogeneity and complexity in hydrological processes, and thus, aiming to identify similar groups of catchments. The input variables (for predicting selected hydrological attributes) are determined by three IVS filter algorithms (partial mutual information, partial correlation input selection, and iterative input selection), then evaluated and compared for among different clusters and when no clustering method is applied. We present and discuss the results for three hydrological attributes – 95% flow percentile (low flows), mean daily discharge (medium flows), and 5% flow percentile (high flows) – in order to account for variability in hydrological conditions, and to investigate the dominant catchment and/or climate characteristics controlling low/medium/high flow predictability at ungauged locations. The findings of our study have implications for reference (donor) catchment selection and hydrological model independent regionalization (aka hydrostatistical or data-driven methods), and are particularly relevant to understanding hydrological similarity and catchment classification in the absence of local runoff observations.