Geostatistics, Fall 2018, Homework 2: Regression

Jonathan Frame
University of Alabama

November 2, 2018

For this assignment I analyzed the predictability of hydrologic signatures from corresponding basin attribute data using three regression methods. A total of 643 basins were included in this analysis. Each regression technique was performed on K-fold cross validation to maintain independence between training and testing data. The results described below come from 21 K-fold groups (20 for training, 1 for testing) with either thirty or thirty-one basins, the difference resulting from unequal division of the full record. The basin data were shuffled randomly before regression analysis to avoid any potential cross-basin correlations, such as geographic proximity.

The first method is multi linear regression using all attributes. This method risks overfitting the linear model to the 37 attributes. After an initial trial I left out hydrologic attribute 13, "high prec timing", because it is constant column of ones, and therefore a linear combination of each other attribute, which can cause errors in the regression equations.

The second method I used was stepwise linear regression. Only mononomial independent variables were included in the analysis with no combinations between variables (such as variable products). An alpha value was set at 0.05 to reject variables which did not add sufficient predictive ability to the multi-linear model. The stepwise algorithm was performed for each hydrologic signature and each K-fold group. The stepwise altorithm resulted in between 5 and 14 attributes used to predict the hydrologic signatures with a mean of about 11.

The third method for regression was principal component analysis. My results showed that it only takes one or two attributes in most cases to explain 95 percent of the variance. To compare the predictability of the method I set the minimum number of components to be equal to those used in Strepwise linear regression.

Figure 1 summarizes the results with bar graphs showing the R-squared values. In general, most hydrologic signatures are best predicted using almost all basin attributes in a multi-linear regression model. In only a few instances is over fitting of concern. Note that results will change slightly when the attached code is run again, due to the random shuffling of data before regression analysis.

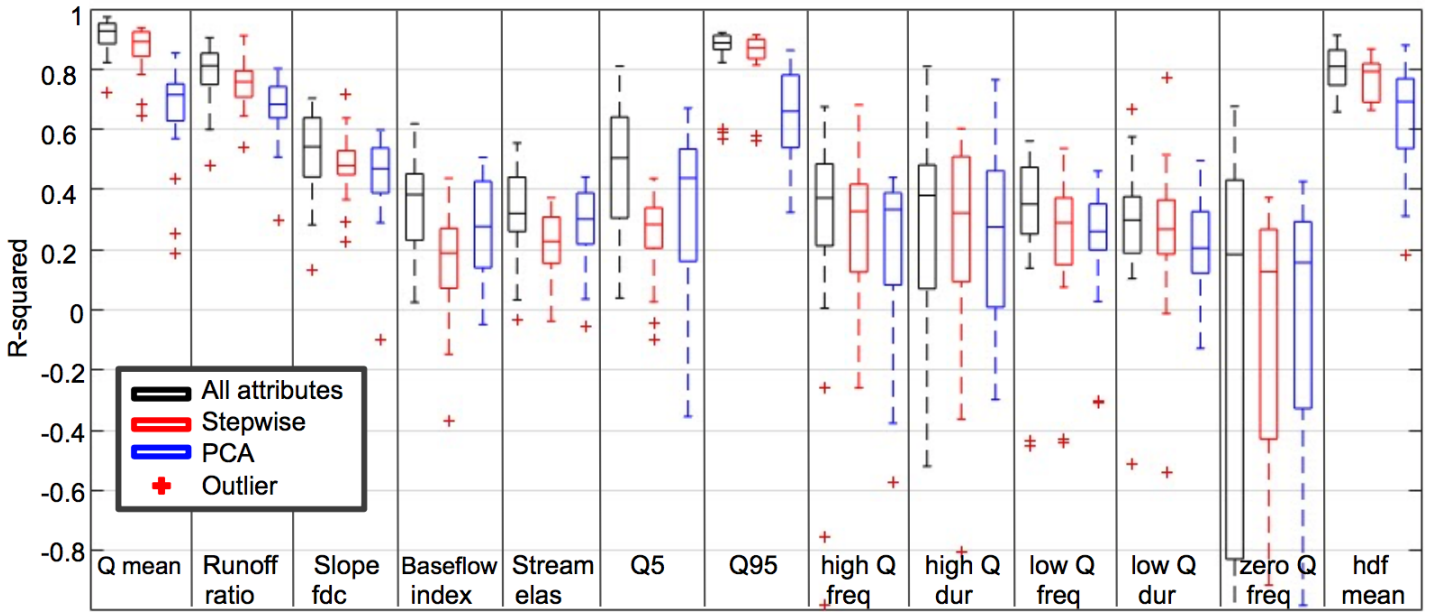

Figure 1: Predictability of hydrologic signatures