Thomas Brower and Dylan Ellis
Data Science Incubator, Project 6
03/11/2020
Presentation Slides

Linear Regression and Baseflow

Introduction

In this project we attempt to predict baseflow values in various rivers using linear regression on a baseflow dataset. The dataset includes the following information regarding the river segment and the surrounding area: the date of measurement, a unique identifier for each segment of a river, the spatial location of the segment, the quantity of water removed by the surrounding area due to evapotranspiration of the adjacent foliage, the quantity of precipitation in the area, the amount of groundwater used for irrigation purposes in the surrounding area in the given month, and the observed baseflow. Linear regression is used to predict the baseflow in a river according to the parameters above. Predicting baseflow is important as it helps us understand our water usage and the load we are placing on our rivers. It is generally acceptable to overdraw rivers in dry months, with the anticipation that the wet months will replenish the lakes that these rivers feed. If the draw on a river exceeds the rate it is replenished at, it can cause environmental catastrophes, like the Aral Sea. We found that irrigation pumping and evapotranspiration were the most predictive explanatory variables. This goes to show how important it is to protect riparian zones and limit irrigation pumping to maintain our rivers.

Dataset

We were given a dataset of observed baseflow values along with various parameters that may influence the baseflow. Baseflow is the measure of flow in a river between precipitation events, in other words, the flow through the river fed by delayed pathways such as springs and mountain run-off maintained independent of the precipitation in the immediate area. Baseflow is an important measure to know if a river is being overdrawn. The attributes included in the dataset are sampling date, segment id, spatial location, evapotranspiration of the surrounding area, precipitation, irrigation pumping, and observed baseflow of the river. For linear regression we only used all of the attributes in the dataset. The dataset was extremely clean and didn't require any cleaning. There were some 0 values, however they did not cause an issue during the analysis.

Analysis Technique

We used linear regression in scikit-learn to predict the baseflow according to the attributes described in the dataset. The predictors were developed in two different ways. First, we performed linear regression on the entire dataset to predict the baseflow for any values of the attributes given to the model. We then separated the data for each river segment and performed linear regression on each segment. This second method will only predict baseflow within a given segment of the river, where the first method can predict for any sample with the desired attributes. Finally, we conducted a grid search to determine which attributes had the most predictive explanatory ability. We only conducted the grid search on 3 attributes:

evapotranspiration, precipitation, and irrigation pumping in the immediate area. We did the search for each individual river segment then reported the average r^2 for each attribute.

After using sk-learn to perform the analysis, we realized that sk-learn doesn't report p-value. We used the same techniques described above but we used statsmodels instead of sk-learn, which allowed us to determine p-values. To avoid p hacking we made sure to consider every attribute that wasn't categorical. The river segments were informative in breaking up the data and computing r squared scores did not change both with and without the river segment data plugged in. Due to this analysis requiring us to run many tests it is sensitive to p hacking. To account for this a Bonferroni correction can be used, where the p-value is divided by the number of tests run prior to determining g whether it is significant. We didn't perform this correction as we were using the p-value as a way to compare the test being run rather than in an attempt to determine the significance of the result.

Results

Linear regression on the entire dataset yielded a low r^2, of 0.2356, and extremely low p-values for all of the parameters, with the highest being 1x10^-2. This means that there was not a strong correlation between the values of the attributes and the observed baseflow.The low p-values tell us that the attributes did have an impact on the baseflow, however they do not fully explain the variance and therefore result in a poor predictor. This can be partially attributed to using different river segments together in developing the predictor. Different river segments exist in different environmental conditions that are not described in the dataset, which can have a huge impact on baseflow. For example, the height of the water table is not considered. A river segment with a high surrounding water table will exhibit higher baseflow than a river with similar attribute values but with a lower water table. This is partially accounted for by the spatial location of the river but that attribute is not sufficient to describe the varying environmental conditions. These conditions are not fully accounted for in the dataset resulting in a poor predictor between river segments.

When breaking the data up into the individual river segments, some of the segments performed extremely well while others others were terrible. For example, segment 123 had a R^2 of 0.9378 and segment 196 had an R^2 of 1.59*10^-5, these values along with several other segments are shown in Table 1. Without knowing more about the segments it is difficult to understand the range of r^2 values. It is likely due to the difference in size of rivers, with small rivers being more sensitive to precipitation than large rivers but having little to no irrigation pumping on them.

Table 1. P-value and R^2 for several segments with relatively high and low values

River Segment	x-coordinate	y-coordinate	Transpiration	Precipitation	Irrigation Pumping	Overall R^2
171	**P-score:** 0.00 **R^2:** 0.002638	**P-score:** 0.00 **R^2:** 0.002638	**P-score:**0.71812 **R^2:** 0.002	**P-score:** 0.00 **R^2:** 0.025051	**P-score:**0.2108 **R^2:** 0.0056	**R^2:**0.095075
152	**P-score:**	**P-score:**	**P-score:**	**P-score:**	**P-score:**	**R^2:**0.4041

	0.177384 **R^2:** 0.328593	0.021055 **R^2:** 0.328593	0.881384 **R^2:** 0.128574	0.000052 **R^2:** 0358144	0.004252 **R^2:** 0.019181	67
53	**P-score:** 0.00 **R^2:** 0.622601	**P-score:** 0.000363 **R^2:** 0.622601	**P-score:** 0.000054 **R^2:** 0.111467	**P-score:** 0.000175 **R^2:** 0.000078	**P-score:** 0.00 **R^2:** 0.010844	**R^2:**0.830648
123	**P-score:** 0.856391 **R^2:** 0.00	**P-score:** 0.856391 **R^2:** 0.00	**P-score:** 0.006283 **R^2:** 0.675050	**P-score:** 0.173983 **R^2:** 0.121063	**P-score:** NaN **R^2:** 0.00	**R^2:**0.937805
159	**P-score:** 0.005524 **R^2:** 0.134058	**P-score:** 0.00 **R^2:** 0.134058	**P-score:** 0.00 **R^2:** 0.223317	**P-score:** 0.00 **R^2:** 0.026936	**P-score:** 0.00 **R^2:** 0.054653	**R^2:**0.522627
196	**P-score:** 0.9584 **R^2:** 0.00	**P-score:** 0.9584 **R^2:** 0.00	**P-score:** NaN **R^2:** 0.00	**P-score:** NaN **R^2:** 0.00	**P-score:** NaN **R^2:** 0.00	**R^2:**0.000016

We found the single most explanatory variable to be the evapotranspiration of the area surrounding the river, as shown in Figure 1. This shows just how important it is to protect our forests and especially riparian zones as they promote groundwater infiltration and allow rivers to maintain higher baseflow. The second most explanatory variable was the amount of water in the surrounding area used for irrigation. This shows how devastating irrigation can be to the flow in a river. The least predictive variables were the x and y spatial location.

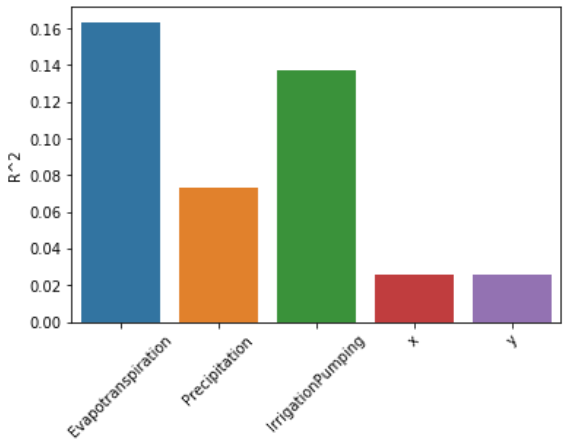

Figure 1. Mean r^2 value over all river segments using the attribute in the x-label as the only attribute for linear regression.

While we were not able to develop a reliable predictor for baseflow, our analysis has shown us some of the important factors that contribute to baseflow. This knowledge can allow us to better manage our waterways to ensure there is ample clean water in the future and to avoid disasters like the Aral Sea.