Suyash Mhetre and Colby Wight
Data Science Incubator CS 5830/6830

Project 6 - Linear Regression in Baseflow Analysis

Introduction

Rivers and streams are typically hotspots for life and vegetation. Many large cities are also centered on these large bodies of water. The size of rivers and streams can vary depending on the season and rainfall. One important attribute of a river is the baseflow. The baseflow indicates what the river's base size is without other factors like rain. It can be especially useful to understand a rivers baseflow especially for farmers and others that depend on a river's water supply to sustain them. We want to know how well we can predict what the baseflow of a river is and what factors contribute most to this.
Slides can be found at:
https://docs.google.com/presentation/d/1cyKZPt13d4ZAtk-2hmMqp_cVqCUi7S484RV-nGCxGjQ/edit#slide=id.g713609f487_0_0

Dataset

For this project we were provided a hydrologic dataset by the instructor called RRCA_baseflow. This contained information on different segments of a river over time. The data was gathered for each river segment approximately once a month. Each of these observations contained the x and y coordinate of the river segment, the evapotranspiration around the river segment, the precipitation, and the irrigation. The baseflow measurement (the dependant variable) was also taken for each observation. No cleaning was necessary for our analysis.

Analysis Technique

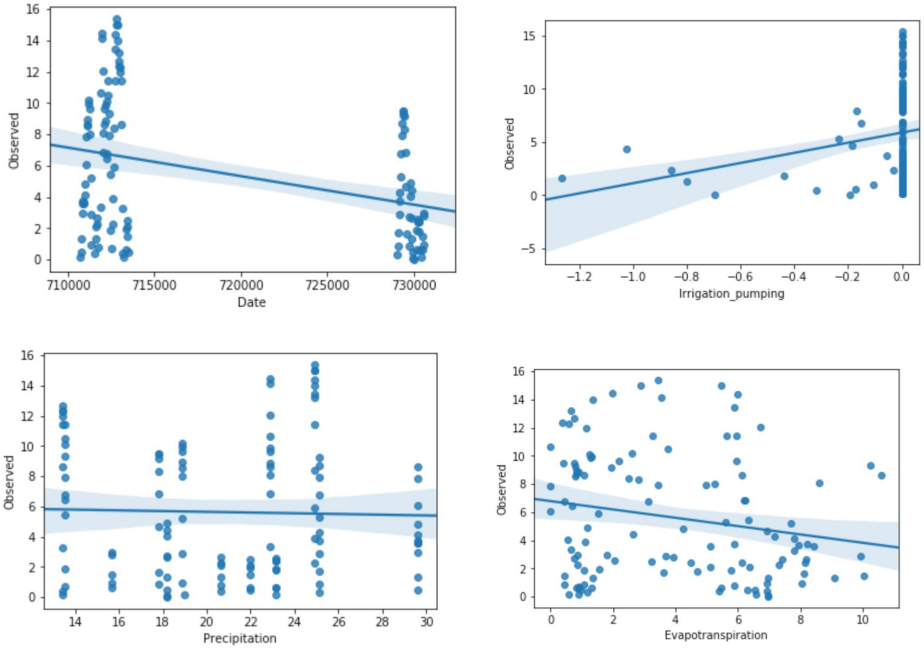
For the analysis we break the problem into two separate ones: one analyzing the baseflow across the whole river and one analyzing the base flow at each given river segment individually. We compare and contrast the results from each approach. For both approaches we use linear regression models for the task of baseflow prediction and variable interpretation.

For the individual segment analysis, we looked at three individual segments, one near the beginning, middle and end of this river. We created a linear regression model using one variable at a time to see how a simple linear regression model would do. We then created a multiple linear regression model using all the variables. We then looked at the model coefficients and p-values to determine which variables to keep in our final model.

We use a similar approach for the whole river analysis but add segment as a variable...

Results

For the Individual segment tests with simple linear regression we found that in general the models that used evapotranspiration and date provided the best R-squared values. However, there was a good amount of variation in the results for the different river segments. We found that in the middle of the river the baseflow decreased in time while at the end of the river the baseflow seemed to increase over time slightly. This seems odd. We include the individual plots for segment 144 below:

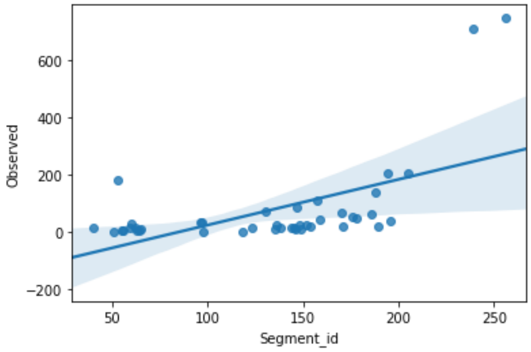


These plots confirm our conclusion in that we see Date and Evapotranspiration help the most in prediction. The line for irrigation pumping looks fairly shar too, but confidence intervals for the coefficient of irrigation pumping are 0.78 to 8.72. Meaning that we can't be sure that the regression line is really that sharp with much certainty.

The multiple linear regression models for individual segments further confirms our findings of evapotranspiration being the most important variable. In our model with all the variables we found the following: evapotranspiration had a coefficient of -0.34 with a p-value of 0.16, Precipitation had a coefficient of -0.062 with a p-value of 0.427, Irrigation pumping had a coefficient of 0.061 with a p-value of 0.98, Date had a coefficient of -0.0002 with a p-value of 0.000. Precipitation and Irrigation had quite large p-values in this case so we decided to leave them out of the final model. Not only did they have large p-values, but the model coefficients were very close to zero indicating that they are not very helpful in this case of predicting baseflow. Another insight we got was that while the p-value for date is highly significant the model coefficient is really close to zero. This tells us that Date is not very helpful in predicting baseflow either, but we will leave it in our final model. We do like the insight we got from evapotraspiration. With a model coefficient of -0.34 we can see that for every unit increase in the scaled evapotranspiration variable we can expect a decrease in the base flow of .33 units.

The final model for the individual segments using evapotranspiration and date achieved a R-squared of .173 for river segment 144. This is not great but much better than using the other variables and we were able to gain some insight into what measurements are important and what are not.

Using segment_Id to find out which segments will be useful for analysis.
Following graph shows that segments within the range of 130 and 200 are significant to use for analysis of individual segments. The R-squared value is 0.3181369878830679.


Observed
Segment_id