Jonah Harmon

Layton Washburn

CS5830 Report 6

https://github.com/JHamoni676/cs5830_project6

https://docs.google.com/presentation/d/1r00vxQV-MYq-rxS7-p2bSZaXIbh_y-TyDQtWEi2d6vA/edit#slide=id.p

1. Introduction

The project utilized the RRCA baseflow data set that recorded measurements such as Evapotranspiration, Precipitation, Irrigation pumping, and the observed baseflow. The date was divided by segment id which represented a specific segment of the river. Each entry was also given a X and Y column that represented the location of the gaging station which was where the observations were recorded. The goal of our project and research were to build a linear regression model that could accurately predict the observed baseflow. This would help farmers be able to better distribute water resources to areas. We removed segmented id's 239 and 256 as they were extreme outliers and skewed our data results. Factors that might have contributed to these segments being so high in baseflow is geographical location, the elevation and slope of the segment and other factors that contribute to a high baseflow. We chose to further break down our data into a season category, which was one of Spring, Summer, Autumn, and Winter. This was important as in our research whether that was googling and reading or talking to students involved in similar research. We found this data to be extremely difficult to work with as the data provided few natural manipulation techniques besides the seasons. In reality there were only three variables that were of any consideration in our research and project which was Evapotranspiration, Precipitation, and Irrigation pumping.

2. Dataset

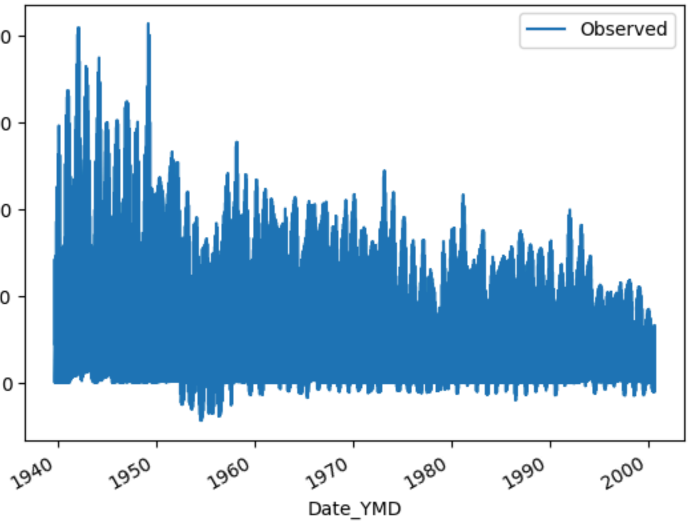
The dataset provided few intuitive ways to split the data into groups to get a better understanding. In our project we decided to break the data up into seasons, one of Spring, Summer, Autumn, or Winter to label a particular entry. This decision was made because we talked to a couple professionals that told us about the varying amounts of baseflow depending on the season. Instances such as farmers using more irrigation water during the hot heat of the summer to water their crops and then easing off the water as the weather gets colder and the crops require less water.

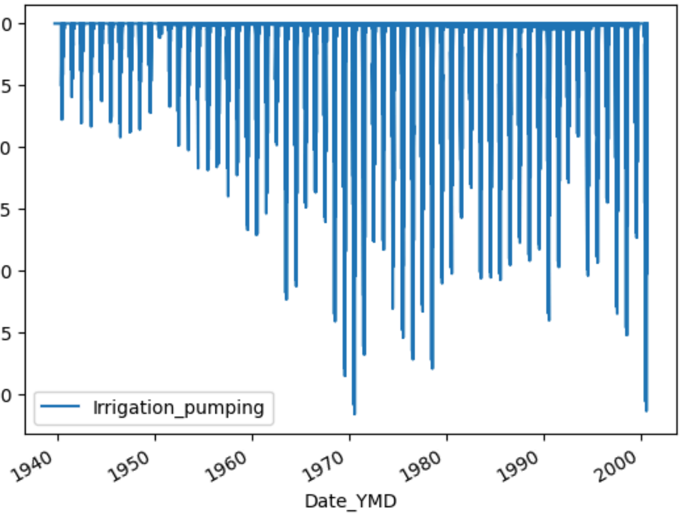
3. Analysis technique

The techniques that we used in our project were removing NA's from the data, removing 2 Segment ids that were outliers, and then splitting each record into a specific seasonal category based on its date. This was necessary because after looking at the graphs and distribution of data, it became clear that there was not an obvious trend in the data except that irrigation pumping had a clear negative correlation with base flow. We tried running a standardization to make the data more consistently formatted but did not see a large impact on our results and so kept the data in the original format. The domain of the data was to be able to create a model to accurately predict baseflow so getting rid of outliers that would skew the data, NA's and other impactful features was necessary to try to feed the model good data.

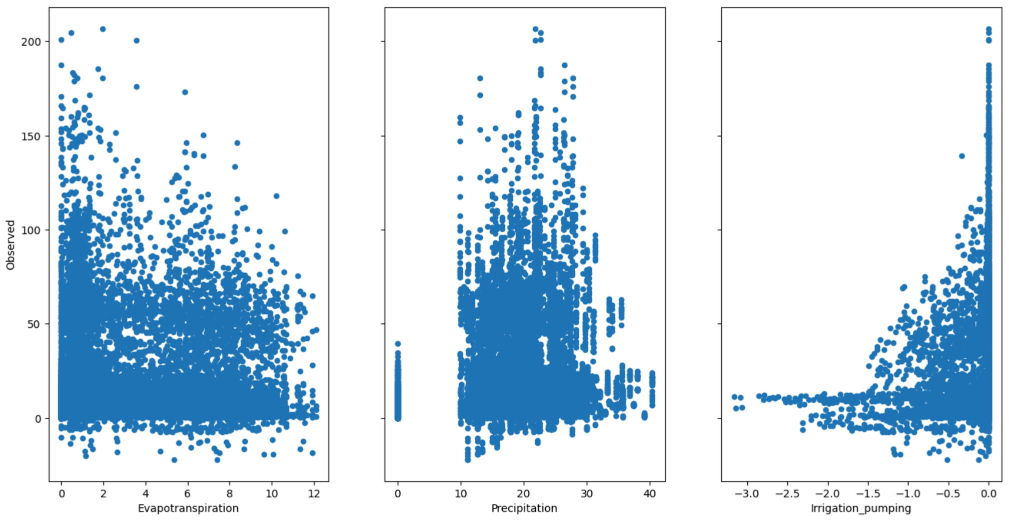
4. Results

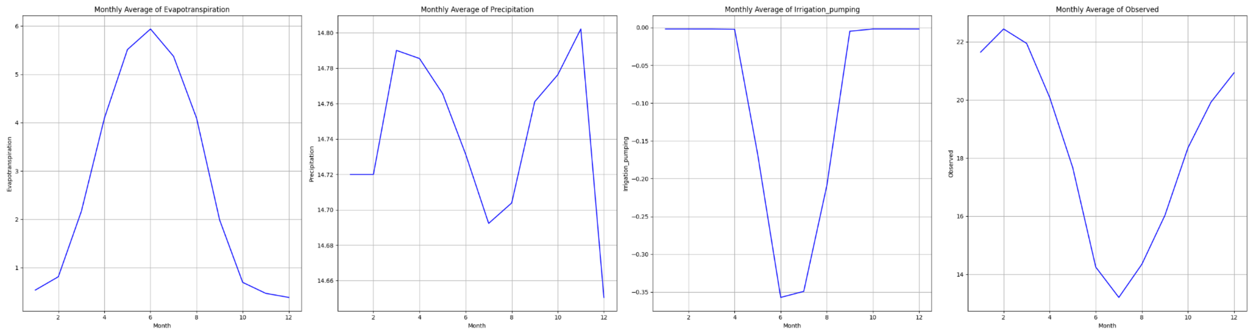
We started off by looking at the general trend of the baseflow and irrigation overtime. We found that the baseflow generally decreased as time increased. We also found that irrigation pumping also has generally increased with time. The observed baseflow over time tells us that factors have been driving the overall baseflow down. We looked into the possibility of dams being built and any droughts that might have happened during those times and didn't find any concrete evidence these outside factors had significant effects.



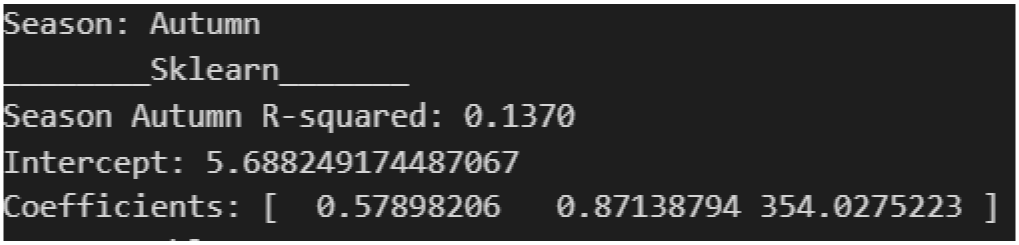


We also found that Evapotranspiration had a slight negative correlation although not strong at all as the trend is a stretch. Precipitation had a similar finding except that it had a small positive correlation, as the precipitation increased the observed baseflow also increased. The irrigation pumping was by far the most convincing as 0 irrigation pumping clearly showed to have a higher baseflow. For this graph, the negative numbers show that water is being pumped out, so the greater the large number, the less base flow will occur.

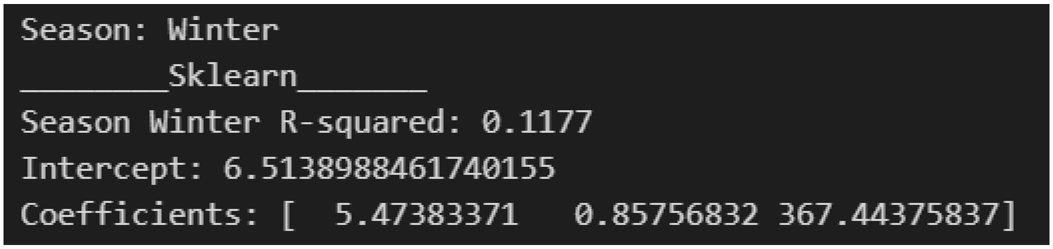




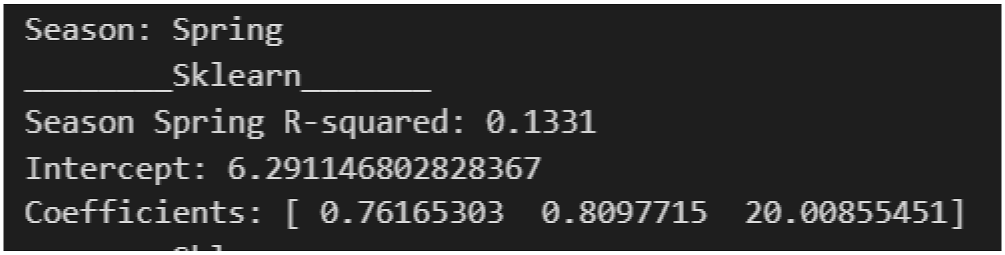
As we divided up the data into seasons, we found that our model had terrible R-values, Autumn 0.137, Winter 0.01177, Spring 0.1331, and Summer 0.1477. This means that our model did not fit the data well at all as closer to 1 means the data fit better. The coefficients followed this format, Evapotranspiration, Precipitation, and Irrigation pumping. What is interesting about this is that Autumn and Winter have very large coefficients for the Irrigation pumping, which means that a unit of irrigation pumping is associated with a 354 or 367 unit increase in observed baseflow which is well above Spring 20 and Summer 7. We talked to a couple of professionals and students who have experience in this area and they talked about how farmers use their irrigation more in the Summer when it is hot which would decrease the amount of base flow. This follows with our data as the seasons with less or decreasing heat have significantly higher irrigation coefficients meaning that a change in the irrigation pumping has a greater effect.



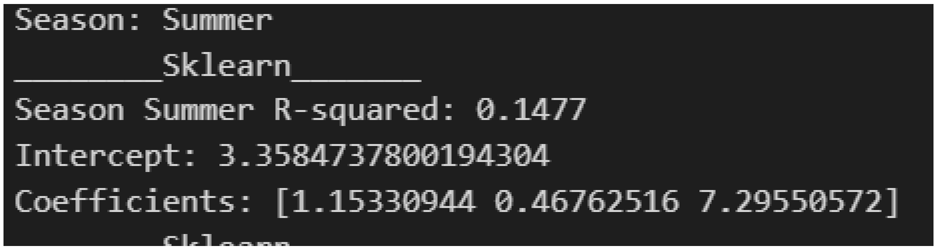
```
Season: Autumn
________Sklearn_______
Season Autumn R-squared: 0.1370
Intercept: 5.688249174487067
Coefficients: [  0.57898206   0.87138794 354.0275223 ]
```



```
Season: Winter
________Sklearn_______
Season Winter R-squared: 0.1177
Intercept: 6.5138988461740155
Coefficients: [  5.47383371   0.85756832 367.44375837]
```



```
Season: Spring
________Sklearn_______
Season Spring R-squared: 0.1331
Intercept: 6.291146802828367
Coefficients: [ 0.76165303  0.8097715  20.00855451]
```



```
Season: Summer
________Sklearn_______
Season Summer R-squared: 0.1477
Intercept: 3.3584737800194304
Coefficients: [1.15330944 0.46762516 7.29550572]
________Sklearn
```

To summarize our findings, this model was not a great predictor of base flow as our model did not fit the data well. This could be for several reasons, first, there were outside factors that we did not take into account or did not find in our research that could have been derived from the data. Second, our data was too grouped, meaning that we might have been able to separate it into smaller groups and label each record to allow the model to more confidently predict base flow. What our model did show was that seasons did have some effect on the base flow as the summer was a time when irrigation pumping would sky rocket causing there to be less base flow. Overall we found that each region based on season, location, and overall need for water in the areas had the most significant effect on baseflow.

5. Technical

This dataset was very difficult to manipulate and we did the best we could. First we tried looking at the predictor columns (Evapotranspiration, Precipitation, and Irrigation Pumping) vs the target (Observed), or baseflow, over each segment. We did this because I learned from a couple people that work for the Department of Agriculture, that each segment of river, especially between damns, is going to be very different in terms of baseflow. That did not give us great linear regression models. So we then attempted to use a KMeans clustering algorithm like was used in the paper linked to the dataset, and that didn’t improve our R Squared values either.

After thinking it through and doing more research, we decided that splitting the data into 4 seasons and making 4 different models would be our best bet. We also removed the outlying segments with extraordinarily high Observed flow (239 and 259). Our p-values in the training sample found that each of the predictors was usually considered to be highly significant. We used Linear regression, scatter plots, and line charts to understand the data as best we could. Part of the data wrangling we did was convert the date in days since 1/1/0000 to a Year-Month-Day format, this made time analyses easier to work with.