CS5830 Project 6 - Linear Regression

Josh Lowe - Kishore Alagarsamy

Introduction

Hydrology is the scientific study of the movement, distribution and quality of water on Earth. It serves as the foundation for environmental management and planning. Efficient management of water resources is important for sustaining the natural ecosystems and supporting human activities. This project focuses on analyzing river flows to better understand them. The objective of this analysis is to leverage statistical techniques to forecast baseflow, enhancing our ability to predict water availability in river basins under varying conditions. These predictions are valuable for water resource management, which informs stakeholders and policymakers about potential changes in water supply which is helpful in planning for use and conservation. Project 6

Our dataset comprises monthly observation of river segments, capturing a range of hydrologic variables including evapotranspiration, precipitation, and irrigation pumping. These factors are crucial for assessing the baseflow, which is the portion of streamflow delayed shallow subsurface flow. By employing multiple linear regressions, we aim to predict the observed baseflow from these predictors. The significance of our findings will be in their potential application because it will give us a predictive tool that can contribute to more informed and effective water management policies. These findings will not be relevant for only environmental scientists and hydrologists but also for government agencies and non-governmental organizations involved in agricultural planning and environment conservation.

Dataset

The dataset employed in this study is a comprehensive collection of hydrological measurements from various river segments, providing detailed monthly observations. Each record captures critical environmental parameters such as evapotranspiration, precipitation, and the amount of groundwater pumped for irrigation. These are all factors that directly influence the baseflow of the adjacent river segments. To ensure robust and an accurate analysis, significant preprocessing was applied to the dataset. Initially, the ‘Date’ column, which initially recorded the number of days since January 1, 0000, underwent normalization to transform this count into a more interpretable format based on the modern calendar starting from January 1, 1900. This adjustment made analyses and comparisons more intuitive. Considering the diverse geographical scope of the dataset, the ‘Segment_id’ attribute, identifying individual river segments, was treated as a categorical variable. To incorporate this into our regression model effectively, one-hot encoding was applied, transforming each segment identifier into a unique binary column, which preserved its categorical nature without imposing assumptions.

The inclusion of variables such as evapotranspiration, precipitation, and irrigation pumping is crucial as they represent the dynamic interaction between human activities and natural processes. Evapotranspiration reflects the sum of water evaporated from the surface and transpired from plants, a key factor in the water cycle influencing river baseflow. Precipitation accounts for the primary natural input of water into the river system, while irrigation pumping represents a significant withdrawal, impacting the natural flow. These attributes form

the foundation of our analysis helping to find the complex interactions that govern river base flow in varying environmental conditions. Preparing the dataset in this way helps provide the basis for applying multiple linear regression, aiming to unveil the relationships between these variables and observed base flows.

Analysis Technique

In this project we used linear regression with feature engineering to model baseflow. Linear regression is suitable because it provides interpretable coefficients, allowing us to quantify the impact of each variable on the baseflow. To capture nonlinear relationships, we created interaction terms and polynomial features for precipitation and irrigation. The use of one-hot encoding for segment identifiers also allowed us to control for regional variations. This approach enabled a more nuanced model while maintaining interpretability, making it suitable for this dataset and domain.

Results

Figure 1: This line plot shows how observed base flow varies over time, providing an overview of seasonal and temporal patterns. By analyzing baseflow over time, we can detect trends, seasonal cycles, or sudden changes. It's essential for understanding how baseflow responds to seasonal changes, such as wet and dry periods.

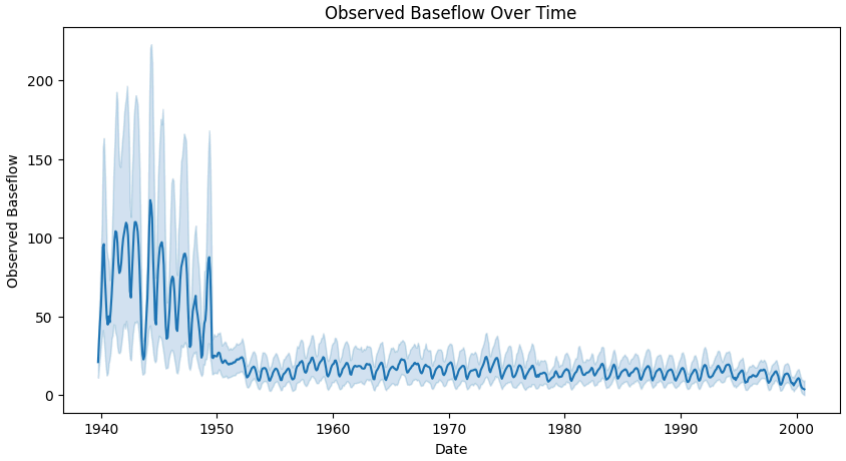

Fig. 1: Observed Baseflow x Time

Figure 2: shows observed baseflow over time with different segments (subregions) highlighted. By adding segment information, we can observe how baseflow varies across different segments, which may experience distinct environmental or land-use factors. This helps to identify variations in baseflow behavior across regions and we can easily address local water management issues differently.

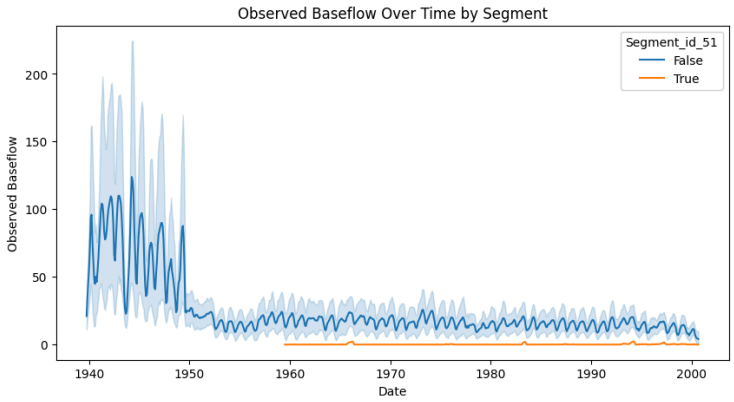

Fig. 2: Observed Baseflow by Segment x Time

Figure 3: This examines the relationship between evapotranspiration and observed baseflow. Evapotranspiration, the sum of evaporation and plant transpiration, often reduces baseflow by depleting groundwater reserves. It's crucial for validating whether increased evapotranspiration, typically during warmer months, correlates with lower baseflow levels.

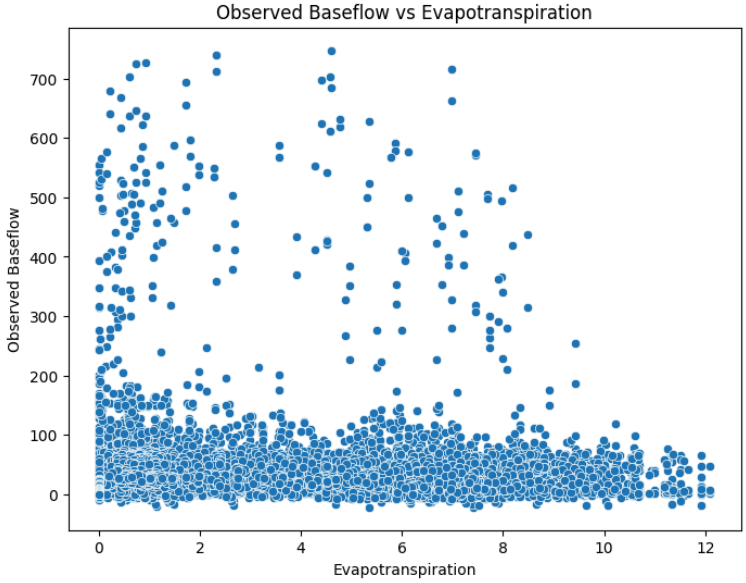

Fig. 3: Observed Baseflow x Evapotranspiration

Figure 4: The relationship between precipitation and observed baseflow. This scatter plot helps confirm whether precipitation directly affects baseflow levels, as expected. It's valuable for demonstrating that increased rainfall contributes to higher baseflow, validating precipitation as a predictor in our model

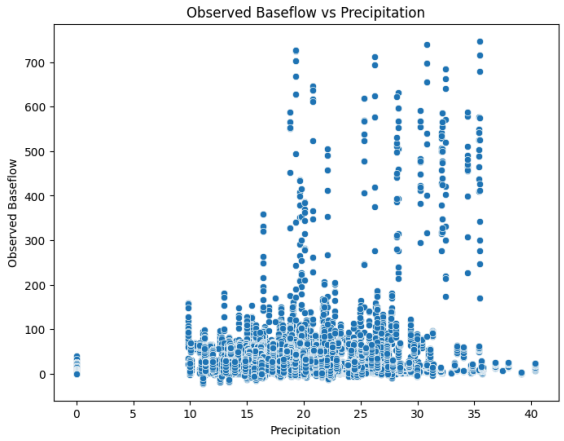

Fig. 4: Observed Baseflow x Precipitation

Figure 5: The relationship between irrigation pumping and baseflow. Irrigation pumping typically draws water from groundwater sources, potentially reducing baseflow. This helps analyze whether increased irrigation pumping is associated with reduced baseflow, making it valuable for understanding how human activities impact natural water resources.

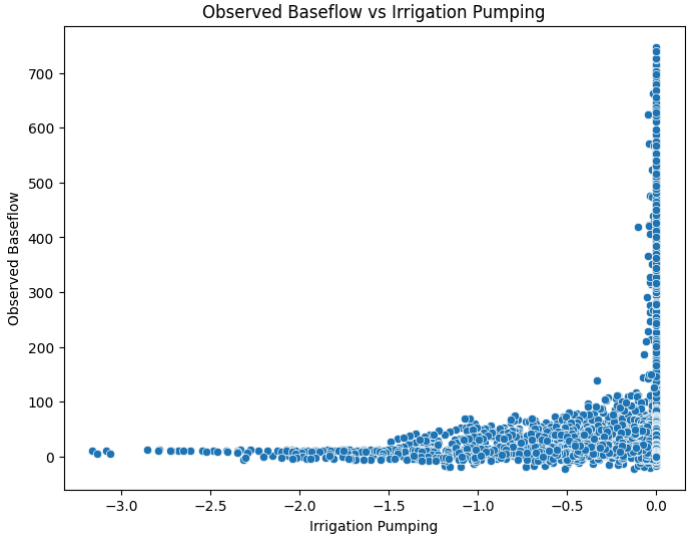

Fig. 5: Observed Baseflow x Irrigation Pumping

Precipitation was found to have a strong positive correlation with baseflow, indicating that increased rainfall significantly contributes to sustaining river flow. Conversely, higher levels of irrigation pumping correlated with decreased baseflow, suggesting groundwater depletion due to agricultural demand. Evapotranspiration also had a negative impact on baseflow, particularly in warmer months. Interaction terms showed that regions with high precipitation and irrigation had unique baseflow patterns, underscoring the need for region-specific water management policies. Overall, the model demonstrated moderate predictive accuracy with a mean absolute error (MAE) of 9.109759217436658, root mean square error (RMSE) of 25.240922657769175, and R-squared (R^2) value of 0.8062687576582308, validating the use of linear regression for this application. These insights are valuable for informing water use policies and predicting seasonal water availability.

Technical

Data Preparation:

Data preparation involved converting the ‘Date’ field into a date-time format, then extracting the year and month to account for temporal trends in the analysis. Segment IDs were transformed into one-hot encoding variables to allow categorical representation in the regression model. Interaction and squared terms for precipitation and irrigation pumping were created to capture potential non-linear variables, providing additional context on how combined factors influenced baseflow.

Analysis:

Linear regression was chosen for its simplicity, interpretability and efficiency in handling datasets with both continuous and categorical predictors. The model can show how each factor individually and in combination (through interaction terms) impacts baseflow, making it well suited for understanding complex, muli-variable relationships in hydrology. Additionally, this method provides straightforward metrics like MAE and R-squared, which help assess the predictive power and fit of the model.

Analysis Process:

Throughout the analysis, adjustments were made to improve model performances, such as including polynomial terms for irrigation and precipitation, which better captured their effects on baseflow. Initial attempts with simple linear terms showed weaker correlations, prompting the addition of interaction terms. Alternative approaches, like ridge and lasso regression, were considered for regularization to prevent overfitting, but initial trials indicated minimal improvements validating the choice of linear regression for this dataset.