Case Study 2

Rocky Mountain River Drainage

Sam Lee & Patric Platts

This study investigates the factors influencing river water flow in the U.S. Rocky Mountains using climate, human activity, and river network characteristics data. We applied Partial Component Regression (PCR) and LASSO Regression to address multicollinearity issues, with LASSO ultimately selected for its balance of predictive performance and interpretability. The LASSO model explained 77.43% of the variance in river flow, with an out-of-sample RMSE of 0.4919. Key factors such as *Precipitation Seasonality* and *Global Stream Order* were identified as significant drivers of river water flow. These results have important implications for water management in the region.

Introduction

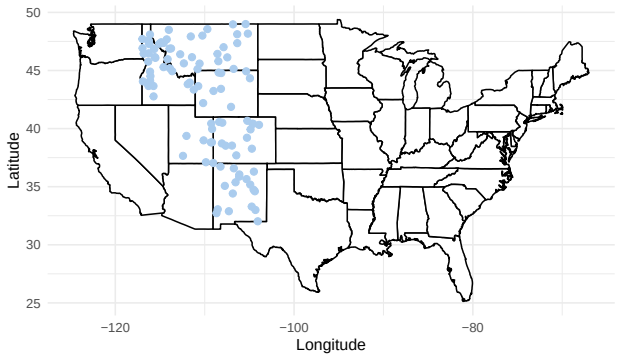
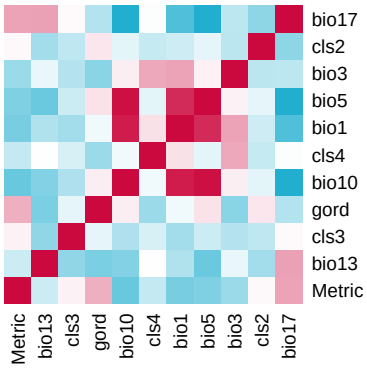


Figure 1: Scatter plot of spatial displacement of each observation of recorded river flow

Ecosystems are shaped by various factors, with rivers playing a critical role in distributing water and nutrients essential for plant and animal life. In the U.S. Rocky Mountains (See Figure 1), the stability of river flow is particularly important, as it influences soil fertility—a key factor for agriculture that sustains both people and livestock. This analysis examines the factors affecting river flow in the Rocky Mountain Region, focusing on human activity, river network characteristics, and climate influences. The data used in this study were collected from multiple rivers in the region to explore how these variables impact overall water flow.

The number of covariates in the data is nearly equal to the number of observations, creating a risk of overfitting due to insufficient local information. In such cases, a standard linear regression model tends to overfitting, capturing noise rather than meaningful patterns, and resulting in poor model performance. High-dimensional data can also lead to "false positives," where unrelated variables appear to be associated with the response variable.

Multicollinearity (see Figure 2) is an issue in the data due to the presence of too many covariates. When multicolinearity is present, the standard errors on our coefficient estimates will tend to be unreliable and hence, the model will have poor generalizability. This issue, combined with the high-dimensional nature of the dataset, necessitates more advanced techniques for model selection and dimensionality reduction such as variable selection and dimensionality reduction techniques. These methods will help identify the most significant factors influencing water flow in rivers throughout the Rocky Mountains.



(a) Figure 2: Correlation matrix of response variable (Metric) and 10 other randomly selected covariates

Methodology

To reduce potential colinearity between the different factors in the data set and arrive at an optimal parsimonious model, we propose two models to assess the overall water flow of water sources in the Rocky Mountains. In this section, we will discuss both candidate models and how these models can be used to answer the research questions at hand.

We first propose a Partial Component Regression (PCR) model. PCR combines Principal Component Analysis (PCA) with linear regression. Under the assumption that the parameters of interest (β) are linear—that is, assuming a one-unit increase in a pth factor (among those we consider) implies a β_p increase in the water flow metric—we leverage this by applying linear regression to the set of orthogonal components computed by PCA[1]. We used ten-fold cross-validation to select the most optimal number of components, k^*; through this process, we chose $k^* = 9$. The strengths with PCR come with its robustness to multicolinearity in the covariate matrix, X. Additionally, PCR performs dimensionality reduction by only selecting the top principal components (in our case, we selected 9) to achieve a parsimonious model.

The tradeoff that comes with using PCR, however, is its lack of interpretability. Since each component is a linear combination of all individual covariates in X, the coefficients derived from our PCR model are not directly interpretable. Additionally, PCR computes and therefore selects components based on the variance of the covariate matrix X, as opposed to each factor's relationship with our response variable, the metric of water flow. Hence, the components may not necessarily contribute to predicting the outcome of interest.

Secondly, we propose fitting a LASSO Linear Regression model to accomplish both dimension reduction through variable selection and interpretability. Similar to our PCR model, we will operate on the assumption that each of our factors have a linear effect on the water flow metric. However, after standardization on the matrix X, LASSO Regression imposes an L_1 penalty[2] to both shrink the estimated coefficients and perform variable selection. Our LASSO Regression model is also suited to handle multicolinearity through the penalization parameter. However, unlike PCR, we can focus on predictive power since there is a direct relationship between the water flow metric and its covariates. Hence, we believe this model to be more interpretable.

When we introduce the penalty parameter, however, the coefficients on this model will be biased. We sacrifice this bias however for a decrease in the variance of the parameters. As a result, to accurately assess

[1]We first orthogonalize the set of all factors of interest, X, through singular-value decomposition, where $X = U\Sigma V'$. We then compute $Z_k = XV_k$, for k number of components where V_k is a subset of V consisting of the first k columns of V. Each column of Z_k is then orthogonal to each other, that is, $Z_i'Z_j = 0\ \forall i \neq j$. Then, performing linear regression, we compute the set of linear γ_k coefficient parameters (where γ_k is of dimension k) using Z_k as the new covariate matrix. Solving for γ_k, $\gamma_k = (Z_k'Z_k)^{-1}Z_k'Y$. $\hat{Y}$ is then computed as $\hat{Y} = Z_k\gamma_k$.

[2]Formally, the L_1 penalty is computed as a vector norm ($||\cdot||$), where, for a vector β with dimension P, $||\beta|| = \sum_{p=1}^{P} |\beta_p|$.

the standard error of each covariate effect, we perform bootstrapping methods to estimate 95% confidence intervals on $\hat{\beta}$.

Model Evaluation

Both models, LASSO and PCR, were evaluated on their in-sample and out-of-sample performance measures.For in-sample evaluation, the adjusted R^2 was used, with LASSO achieving 0.7318 and PCR achieving 0.6549, suggesting that LASSO only provides a marginally better fit to the data.Out-of-sample prediction performance was assessed using the root mean square error (RMSE), where LASSO recorded an RMSE of 0.4919 and PCR 0.5292. While both models showed similar predictive ability, LASSO was ultimately selected due to its superior interpretability.

PCR reduces dimensionality by grouping variables into components, which explain portions of variability, but the complexity of interpreting these components—where variables contribute with varying weights—makes it challenging to discern the impact of individual variables on river water flow. In contrast, LASSO enhances interpretability by shrinking the coefficients of less important variables to zero, thus retaining only the key variables. The decision to use LASSO was driven by its balance between reasonable predictive performance and straightforward interpretability, making it a more practical choice for this analysis. Our LASSO model can be represented by the general solution to the minimization problem shown in Equation 1:

$$\arg\min_{\beta} \sum_{i=1}^{n} (y_i - x_i'\beta)^2 + \lambda \sum_{p=1}^{P} |\beta_p| \quad (1)$$

$$\hat{Y} = X\hat{\beta} + \lambda^* \sum_{p=1}^{P} |\hat{\beta}_p| \quad (2)$$

We make predictions with our model using Equation 2 after finding the solution set to Equation 1, where λ^* is the penalty parameter found using 10-fold cross validation. We note that originally, β and $\hat{\beta}$ are of dimension P, where P is the number of total covariates we include in our model. For a full description of each covariate we refer the interested reader to the data description source here.

The model makes several assumptions, including linearity and independence. It assumes a linear relationship between the predictors and the outcome, which was assessed through exploratory data analysis. The independence assumption holds that the data for each river is collected independently, focusing on one observation at a time.

Results

With our selected model, we estimated the standard errors through bootstrapping[3] to assess the the 95% confidence intervals on $\hat{\beta}$. These results are summarized in Table 1.

Table 1 lists and describes the most significant climate, river network, and human factors that impact overall river flow. Of the factors that are most significant are *Precipitation Seasonality*, *Mean Somewhat Excessive Drainage Class*, and *Global Stream Order*. Since these coefficients are linear, they can be interpreted as such:

[3]To estimate the standard errors of $\hat{\beta}$, we first computed $B = 10,000$ bootstrap samples from Y (the water flow metric) and our covariate matrix X with replacement of size $K = N = 100$ where N was the total number of observations in the data set. Using the optimal penalty parameter, λ^*, as computed through our cross-validation step previously, we estimated B # of LASSO Regression models and computed the standard error of each $\hat{\beta}_j$ for $j = 1, ..., P$ given our $\hat{\beta}$ vector of dimension P through the following computational sequence: (1) For each $\hat{\beta}_j$, compute $\bar{\hat{\beta}}_j = \frac{1}{B}\sum_{b=1}^{B} \hat{\beta}_j$. (2) $SE(\hat{\beta}_j) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\beta}_j - \bar{\hat{\beta}}_j)^2}$. (3) We compute the 95% C.I. as $\left(2\hat{\beta}_j - \hat{\beta}_{j_{\text{boot}}}^{(0.975)}, 2\hat{\beta}_j + \hat{\beta}_{j_{\text{boot}}}^{(0.025)}\right)$, where $\hat{\beta}_{j_{\text{boot}}}^{(0.975)}$ and $\hat{\beta}_{j_{\text{boot}}}^{(0.025)}$ are the 97.5th and 2.5th quantiles of the bootstrapped distributions of $\hat{\beta}_j$, respectively.

Table 1: Lasso Coefficient Estimates

Covariate	Description	Estimate	95% CI	Inclusion Frequency
(Intercept)		0.125*	(0.02, 0.207)	1.000
bio10	Mean Temperature of Warmest Quarter (degrees Celsius)	-0.032	(-0.063, 0.067)	0.261
bio15	Precipitation Seasonality (Coefficient of Variation) (milimeter)	-0.16*	(-0.313, -0.034)	0.981
bio18	Precipitation of Warmest Quarter (milimeter)	-0.039	(-0.077, 0.086)	0.559
cls1	Evergreen_Dec_Needle_Trees (percent)	0.062	(-0.175, 0.124)	0.730
cls2	Evergreen_Broadleaf (percent)	0.083	(-0.162, 0.166)	0.958
cls5	Shrubs (percent)	-0.038	(-0.075, 0.145)	0.553
cls8	Regularly Flooded Vegetation (percent)	0.106*	(0.046, 0.211)	0.904
CumPrec03	Cumulative March Precipitation for the Watershed Upstream of Grdc Station (milimeter)	0.031	(-0.161, 0.063)	0.467
CumPrec04	Cumulative April Precipitation for the Watershed Upstream of Grdc Station (milimeter)	0.129*	(0.035, 0.257)	0.416
CumPrec05	Cumulative May Precipitation for the Watershed Upstream of Grdc Station (milimeter)	0.024	(-0.112, 0.048)	0.273
gord	Global Stream Order from Stream Dem (Predicted Relationship with Area) (categorical)	0.215*	(0.12, 0.43)	0.974
Lon	Longitude	-0.177*	(-0.354, -0.064)	0.795
meanPercentDC_ModeratelyWell	Mean Moderately Well Drained Soil (percent)	-0.028	(-0.056, 0.056)	0.541
meanPercentDC_Poor	Mean Poorly Drained Class (percent)	0.072	(-0.024, 0.144)	0.743
meanPercentDC_SomewhatExcessive	Mean Somewhat Excessive Drainage Class (percent)	0.165*	(0.012, 0.33)	0.944
MeanPrec07	Mean July Precipitation for the Watershed Upstream of Grdc Station (milimeter)	-0.016	(-0.032, 0.202)	0.579

Hence, a one unit increase in the coefficient variation of the (scaled) *Precipitation Seasonality* decreases the water flow by -0.188, on average.

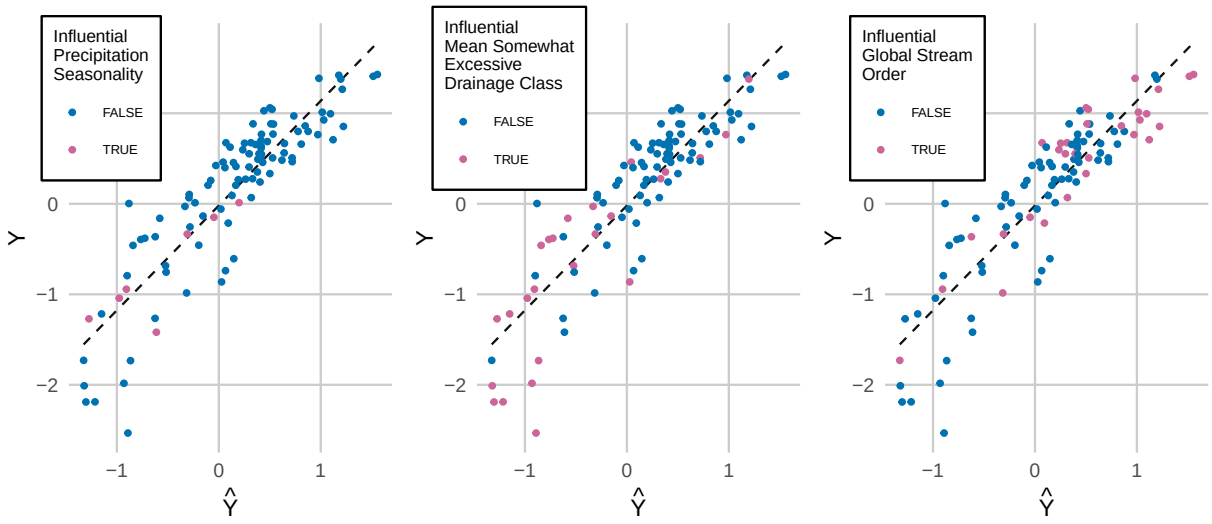


Figure 3: A comparison of actual and predicted values using LASSO Regression

We visually summarize the most significant effects[4] in Figure 3. For an exact fit, $\hat{Y} = Y$, and hence, for a given factor, he closer an observation is to the equilibrium line, we say the more *influence* that factor had in predicting the water flow metric of that observation. Under this pretext, we acknowledge the large variance in the *Mean Somewhat Excessive Drainage Class*. While several factors may be confounding a single observation, given that the factors above are statistically significant, it is likely that the observations close to the 45-degree line are well predicted by these influential factors. Table 1 also records the *inclusion frequency*[5]. This is a metric of robustness to variation in random sampling. Hence, a larger inclusion frequency indicates a stronger dependency with water flow. We use inclusion frequency in part to assess how well these factors explain overall flow. We point out here that *Precipitation Seasonality* has the highest inclusion frequency.

Through the fitted LASSO Regression model, 77.43% of the variance in the water flow metric can be explained by our selected covariates. When corrected by the number of factors, we obtain an adjusted R-squared of 73.18%. We believe that this reflects the parsimonious fit of our selected model. Using leave-one-out-cross validation (LOOCV), we computed an out-of-sample RMSE of 0.4919. Thus, on average the out-of-sample prediction is 0.4919 away from the actual metric of river flow.

Conclusion

The goals of this study were to model the factors that influence river water flow in the U.S. Rocky Mountains, accounting for potential multicollinearity in the data set. Two candidate models were proposed—PCR and LASSO Regression—to handle the high-dimensional data. While PCR effectively reduces dimensionality, LASSO was chosen for its clearer interpretability and ability to select significant covariates. The LASSO model revealed key factors like *Precipitation Seasonality* and *Global Stream Order* as significant drivers of water flow. Despite the model's utility, limitations include the bias introduced by LASSO's penalty parameter and the potential for over-simplification of complex environmental interactions. We also acknowledge the model's shortcomings when it comes to spatial correlation (see Figure 1) in the data structure. Future studies could address this spatial correlation and investigate non-linear models or explore the effects of additional ecological and anthropocentric variables. These next steps would enhance our understanding of water flow dynamics in the Rocky Mountain region and improve water management strategies.

Teamwork

We worked on the coding and analysis together (git contributions can be viewed here), and then we split up writing the report: Sam wrote up the Proposed Methods, Results; Patric wrote up the Methodology, and the Introduction. We completed the Abstract and Conclusion together.

[4] For a given jth factor, influential effects are classified as all observations in the set, $\{x_{ij} : x_{ij} \leq X_j^{(0.05)} \text{ or } x_{ij} \geq X_j^{(0.95)}\}, i = 1, ..., N$.

[5] For a given jth factor, using the bootstrap distributions, we calculate the inclusion frequency as $\frac{1}{B}\sum_{b=1}^{B} \mathbb{1}(\hat{\beta}_j^{\text{lasso, } b} \neq 0)$.