Case Study 2

Rocky Mountain River Drainage

Sam Lee & Patric Platts

Abstract

Introduction

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, 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.

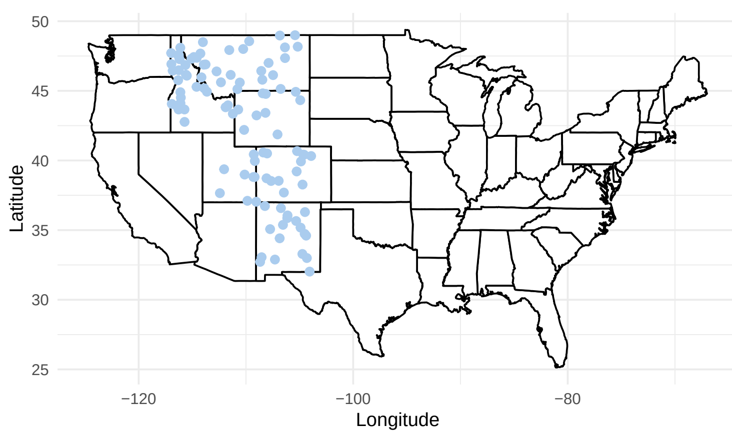

Figure 1: Scatter plot of spatial displacement of each observation of recorded river flow

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. To address these issues, this analysis applies variable selection and dimensionality reduction techniques. These methods will help identify the most significant factors influencing water flow in rivers throughout the Rocky Mountains.

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

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 the standard error of each covariate effect, we perform bootstrapping methods to estimate 95% confidence intervals on β.

---

[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, $||\beta|| = \sum_{p=1}^{P} |\beta_p|$.

Model Evaluation

Both models, LASSO and PCR, were evaluated using in-sample and out-of-sample performance measures. To assess the in-sample fit, the adjusted R^2 was used. LASSO achieved an adjusted R^2 of [insert value], while PCR achieved an adjusted R^2 of [insert value]. This indicates that [LASSO/PCR] provides a better in-sample fit to the data.

For out-of-sample prediction performance, the root mean square error (RMSE) was compared between the two models. LASSO had an RMSE of [insert value], while PCR had an RMSE of [insert value]. Although both models yielded similar RMSE values, indicating that either method could be appropriate for analyzing this dataset, LASSO was ultimately chosen due to its interpretability.

PCR reduces dimensionality by grouping variables into components, each explaining a portion of the variability. However, the complexity of interpreting these components—since different variables have varying weights within each component—makes it difficult to determine the individual contribution of each factor to river water flow. In contrast, LASSO simplifies interpretation by shrinking less important variables to zero, leaving only the key factors that directly impact water flow.

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

Talk about the assumptions of the model cleary

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*.

[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 $j = 1, \dots, P$ 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*	(0.017, 0.208)	1.000
bio10	Mean Temperature of Warmest Quarter (degrees Celsius)	0	(0, 0.102)	0.207
bio15	Precipitation Seasonality (Coefficient of Variation) (milimeter)	-0.188*	(-0.346, -0.079)	0.990
bio18	Precipitation of Warmest Quarter (milimeter)	-0.014	(-0.028, 0.113)	0.482
cls1	Evergreen_Dec_Needle_Trees (percent)	0.034	(-0.205, 0.068)	0.664
cls2	Evergreen_Broadleaf (percent)	0.082	(-0.126, 0.163)	0.934
cls5	Shrubs (percent)	-0.019	(-0.037, 0.16)	0.478
cls8	Regularly Flooded Vegetation (percent)	0.081*	(0.022, 0.161)	0.846
CumPrec03	Cumulative March Precipitation for the Watershed Upstream of Grdc Station (milimeter)	0.049	(-0.157, 0.097)	0.546
CumPrec04	Cumulative April Precipitation for the Watershed Upstream of Grdc Station (milimeter)	0.132*	(0.029, 0.264)	0.494
gord	Global Stream Order from Stream Dem (Predicted Relationship with Area) (categorical)	0.174*	(0.071, 0.348)	0.950
Lon	Longitude	-0.177*	(-0.354, -0.096)	0.750
meanPercentDC_Poor	Mean Poorly Drained Class (percent)	0.03	(-0.06, 0.059)	0.482
meanPercentDC_SomewhatExcessive	Mean Somewhat Excessive Drainage Class (percent)	0.18*	(0.046, 0.359)	0.946
MeanPrec07	Mean July Precipitation for the Watershed Upstream of Grdc Station (milimeter)	-0.002	(-0.003, 0.208)	0.536
MeanTemp05	Mean May Temperature (degrees Celsius)	-0.002	(-0.005, 0.083)	0.144

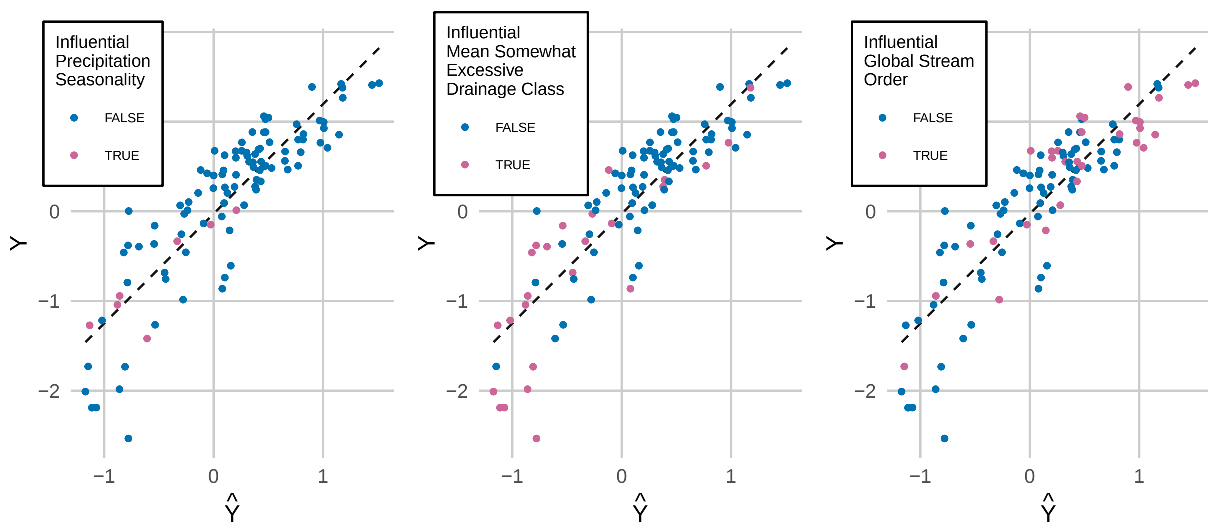

Figure 2: A comparison of actual and predicted values using Lasso Regression

We visually summarize the most significant effects[4] in Figure 2. For an exact fit, $\hat{Y} = Y$, and hence,

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

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*. 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, 73.93% 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 69.39%. 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 0.5118. Thus, on average the out-of-sample prediction is 0.5118 away from the actual metric of river flow.

Conclusion

$X_j^{(0.95)}\}, i = 1, \ldots, N$.

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