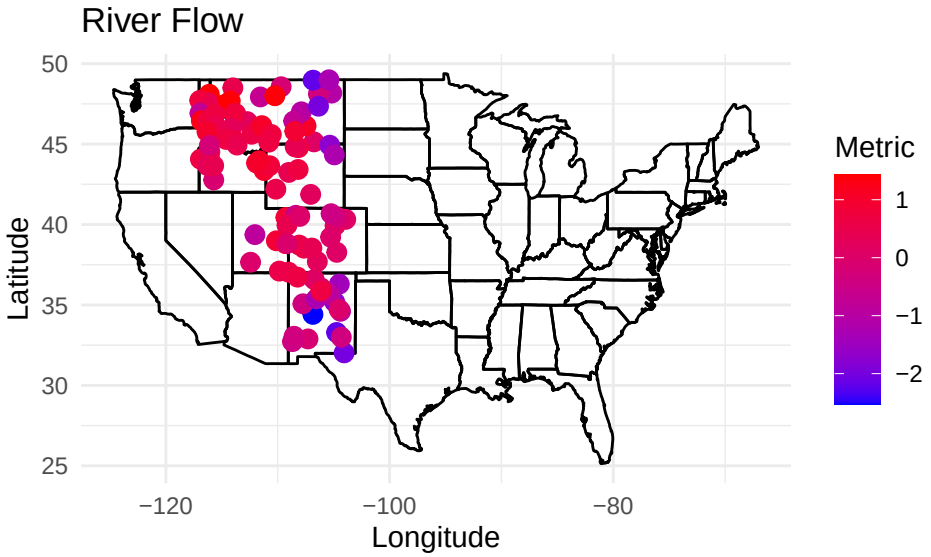
STAT 536 HW 1 Report

Ty Hawkes and Brigg Trendler

2024-09-30

Abstract

1. Introduction: Problem Statement and Understanding



Rivers are responsible for carrying water and nutrients to various parts of the earth. Many animals, including humans, rely on rivers and the life they bring to thrive in otherwise uninhabitable areas. In this project, we work to analyze how various factors (human, river network, and climate) impact the overall river flow in the Rocky Mountain area.

Through an initial exploratory data analysis, we found that we have 95 covariates and 102 observations. This means, should we create a linear regression model with every variable, the model would likely be severely overfit, which would lead to several potential problems, including multicollinearity, high variance in coefficients, and interpretability issues.

The data set also contains monthly averages for precipitation, temperature, and cumulative precipitation, as well as annual averages for those same variables. This is potentially problematic because including both the yearly and monthly averages would introduce multicollinearity into the model. Furthermore, some of the variables are direct linear combinations of others. For example, the variable bio7 gives the annual temperature range, given by subtracting the max temperature from the minimum, both of which are stored as bio5 and bio6. Thus we want to avoid including these in our model simultaneously.

In this project, we hope to find a predictive model for river flow based on the given variables while simultaneously avoiding incorporating the aforementioned problems into our model.

2. Methodology

2.1 Proposed Methods

To combat the high dimensionality of our data set, we propose two models: a linear regression model constructed from a LASSO regularization, and a linear regression model created from a bi-directional selection algorithm using the AIC as a scoring metric.

LASSO regularization helps reduce dimensionality selecting the most relevant predictors. It improves interpretability by producing a sparse model and reduces overfitting. However, it can introduce bias into coefficient estimates, and can struggle when trying to handle highly correlated variables. In the context of the problem of this case study, a model built by LASSO would solve the dimensionality problem, however the colinearity between variables might pose a problem for this method.

A forward and backward selection algorithm balances model fit and complexity. It also automates variable selection and has the advantage of navigating correlated predictors better than a pure forward or pure backward selection algorithm. Unfortunately, it is a greedy algorithm, meaning it may not find the best model. It also does not handle multicollinearity well and can be computationally expensive. A forward and backward selection model in the context of this problem would also solve the dimensionality issue.

In both cases, we will need to check the classic LINE assumptions when using linear regression: linearity, independence, normally distributed errors, and equal variance.

2.2 Model Evaluation

3. Results

3.1 What are the biggest climate / river network / human factors that impact overall river flow?

The table below shows the beta coefficients of the factors that remained after variable selection. Most variables in our model had a significant impact on overall river flow.

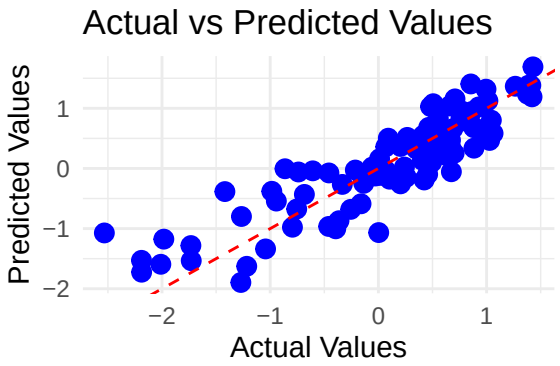
Term	Category	Estimate	Std. Error	Pr(>
(Intercept)		103.18286	22.61833	1.61e-05 ***
bio15	Climate	-0.59865	0.13978	4.65e-05 ***
CumPrec05	Climate	2.20575	0.63710	0.000826 ***
mPDC_SomewhatExcessive	Land Cover	0.42028	0.16888	0.014682 *
bio18	Climate	0.73360	0.37259	0.052074 .
cls2	Land Cover	317.43807	65.55078	5.37e-06 ***
Lon		-0.08278	0.02067	0.000128 ***
cls9	Land Cover	-1.32608	0.34508	0.000228 ***
bio14	Climate	-1.25629	0.39629	0.002090 **
gord	Network	0.44521	0.11739	0.000271 ***
mPDC_ModeratelyWell	Land Cover	-0.19627	0.09066	0.033074 *
cls8	Land Cover	2.17050	1.09753	0.051066 .
cls1	Land Cover	0.17930	0.09525	0.063035 .

Each category had at least one significant factor affecting river flow. Climate factors included bio15 (Precipitation Seasonality (Coefficient of Variation)), CumPrec05 (Cumulative May precipitation for the watershed upstream of grdc station), and bio14 (Precipitation of Driest Month); Land cover factors included meanPercentDC_SomewhatExcessive (mean somewhat excessive drainage class), cls2 (Evergreen_Broadleaf), cls9 (Urban), meanPercentDC_ModeratelyWell (mean moderately well drained soil); and network factors included gord (global stream order from stream dem (Predicted relationship with area)).

3.2 How well do these factors explain overall flow?

Our model does a decent job explaining the overall flow of rivers. With these selected co-variates, our model explains nearly 82% of the in-sample variance as measured by the r-squared value.

3.3 How predictive of overall flow are these identified factors?



We estimate that our model will do a reasonable job predicting overall flow on our-of-sample data. As demonstrated in the figure above, using a leave-one-out cross validation lead to an RMSE of .4077, a significant improvement over the base intercept model with an RMSE of .8845. This reduction in error indicates that our model is better equipped to capture the variability in river flow than taking the average river flow of each river.

4. Conclusions

Teamwork