STAT 536 HW 1 Report

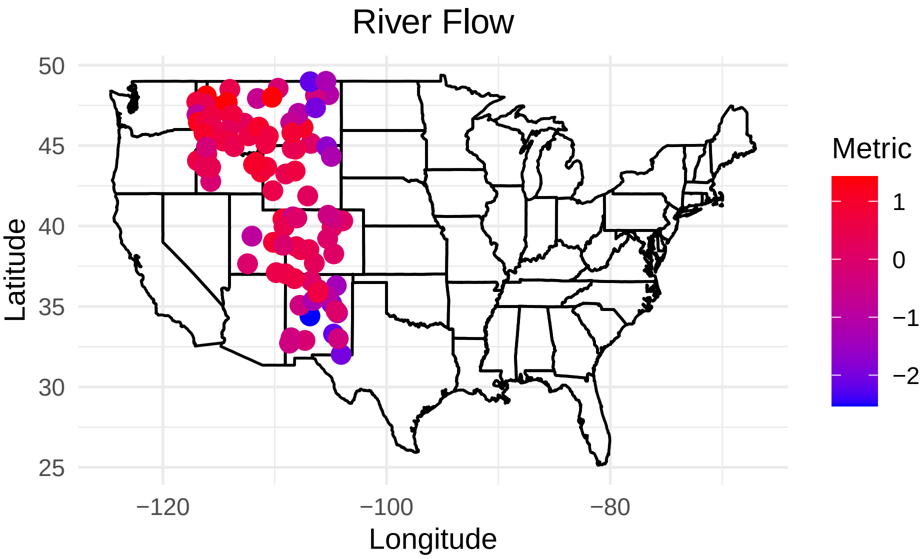
Ty Hawkes and Brigg Trendler

2024-09-30

Abstract

This report investigates the impact of various factors on river flow in the Rocky Mountain region, utilizing a dataset comprising 95 covariates and 102 observations. Given the potential for overfitting due to high dimensionality and multicollinearity, we implement two modeling approaches: LASSO regularization and bi-directional variable selection based on Akaike Information Criterion (AIC). Through exploratory data analysis, we identify significant climate, land cover, and network variables affecting river flow. Our final model explains approximately 82% of the variance in river flow, demonstrating its predictive capability with a root mean square error (RMSE) of 0.4077 in cross-validation, significantly outperforming a baseline intercept model. We conclude that while our model effectively captures the variability in river flow, further research with diverse datasets is necessary to enhance its predictive performance and interpretability.

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

Both models seemed to take a conservative approach to interpretability by selecting the key components and favoring model simplicity. We decided to test both models using leave-one-out cross validation and found that the forward-and-backward selection algorithm model appeared to perform better, both in sample and out of sample.

Model	R^2	RMSE
LASSO	0.8065	0.484
Forward and Backward Selection	0.8297	0.4077

Our Final Model:

$$\begin{aligned}\widehat{\text{Flow}}_i = 103.18286 &- 0.59865 \times \text{bio15}_i + 2.20575 \times \text{CumPrec05}_i + 0.42028 \times \text{mPDC}_i \\ &+0.73360 \times \text{bio18}_i + 317.43807 \times \text{cls2}_i - 0.08278 \times \text{Lon}_i - 1.32608 \times \text{cls9}_i \\ &-1.25629 \times \text{bio14}_i + 0.44521 \times \text{gord}_i - 0.19627 \times \text{mPDC}_i \\ &+2.17050 \times \text{cls8}_i + 0.17930 \times \text{cls1}_i\end{aligned}$$

To verify our model is operating under the correct assumptions, we plotted the residuals vs fitted values graph and found that the residuals are randomly scattered around zero, indicating that has mostly captured the true relationship between the predictors and the river flow metric. This also indicates that the assumptions of linearity and constant variance are likely met.

We generated a normal Q-Q plot and found that the points fall close to the diagonal line, which suggests that the data likely follows from a normal distribution, again validating one of our underlying assumptions when working with linear regression models.

The scale-location plot reveals a downward trend of our standardized residuals. This indicates that there is a bit of heteroscedasticity in the data. However, the part that appears to be heteroscedastic also has the fewest data points. The bulk of the residuals appear to be clustered around 0, which may indicate that the data is homoscedastic enough for our purposes.

We also checked the residuals vs leverage and find that there are no blatant problems. There are no distinct patterns, meaning the model fits well even in the presence of the outliers in the data set.

Finally, we made sure our final model did not contain colinear variables (such as the yearly averages and bio7) as the information contained in those variables were also found in other variables or was deemed not as important by the feature selection algorithm.

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.

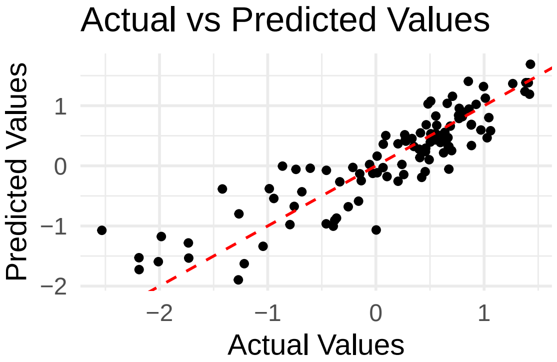
Term	Category	Estimate	Std. Error	Pr(>
(Intercept)		103.18286	22.61833	1.61e-05 ***
bio15	Climate	-0.59865	0.13978	4.65e-05 ***
CumPrec05	Climate	2.20575	0.63710	0.000826 ***
mPDC_SomewhatExcessive	Land Cover	0.42028	0.16888	0.014682 *
bio18	Climate	0.73360	0.37259	0.052074 .
cls2	Land Cover	317.43807	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. This means that only 13% of the factors in the data are needed to explain 82% of the variation in river flow.

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

We used a linear model to evaluate the impact that different factors have on river flow. We found that despite there being 95 different variables available to us, over 82% of the variation in river flow can be explained by 12 variables alone. At least one climate, land cover, and network variable had a significant influence on the overall river flow. Using cross validation, we were able to show that our model had the capability of predicting river flow out of sample much better than simply taking the average river flow across all rivers.

The biggest shortcoming of our model is it's ability to predict. To better understand the factors that contribute to river flow, we had to reduce the number of variables and avoid transformations that could have reduced its interpretability. Additionally, our model was trained on a geo-spatially correlated area of land in the united states, introducing unaccounted for violations of the independence assumption.

In the future, we would like to collect more data from different parts of the world to get a better idea of how strong our predictors actually are. Additionally, we would like to make two models: one to interpret coefficients, and another to make excellent predictions. We would also like to account for geo-spatial correlation in these models.

Teamwork

Ty coded up potential models for the data, checked assumptions, helped troubleshoot code, and performed cross validation on them. Ty also wrote the results, conclusions, and the plots in the report

Brigg also coded up potential models for the data, helped troubleshoot code, and tediously wrote long formulas in R. Brigg also wrote the introduction, methodology, and abstract.