Rocky Mountain River Drainage

Emily Liu and Josh Christensen

10/4/2021

Abstract

Rivers play an important role in our earth's ecosystem. Rivers provide food and habitat for plants and animals and are a primary source of water for crops. There are various rivers found in the Rocky Mountain region, each with its own geographical setting and water flow patterns. Knowing what factors impact overall water flow in these areas may help farmers make better agricultural decisions to utilize the maximum amount of river flow in each area. In this analysis, we use a generalized boosted regression model to investigate various factors such as human influences, river network and climate that may impact the overall river flow of various rivers in the Rocky Mountain region. Through our modeling, we also quanitfy how predictive these factors are for overall flow. In doing so, we hope to help farmers better understand the causes of water flow to increase the effectiveness of crop irrigation systems and produce more productive farmlands.

Exploratory Data Analysis

In our dataset there are 102 observations with ninety-eight different explanatory variables that could affect overall water flow. These variables include average temperature, average precipitation, environmental surroundings, human regional population, soil drainage, and other variables. Through a pruned decision tree, we found that some of the most important variables that could affect river water flow are seasonal variation in precipitation and cumulative precipitation. Below is also a distribution of the response variable, metric, and we observe that the distribution is left-skewed with a positive median.

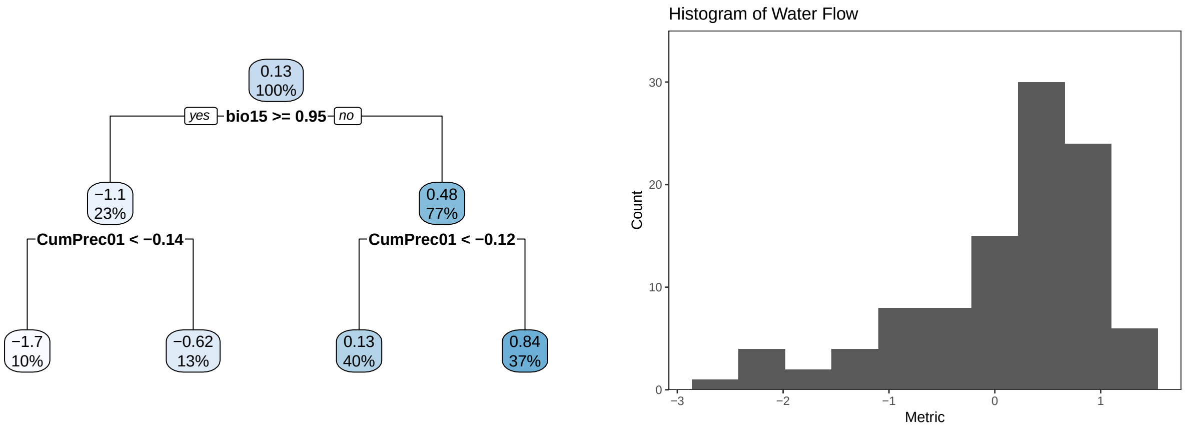

Figure 1: Pruned decision tree and response variable distribution.

Futhering our exploratory analysis, we noticed that in this data set that there are almost the same number of observations as explanatory variables. This would create issues in linear regression models because there is too much flexibility for the model. It has so many explanatory variables it can match each observation almost perfectly, leading to overfitting and imprecise coefficients. Therefore, a model that accounts for this type of dataset where the number of explanatory variables and observations are almost the same, will need to be used in order to produce accurate and reliable results. We detail such a model in the following section.

Methods and Models

As seen above, we began our transition from exploratory data analysis to modeling with a regression tree. The tree was created using recursive binary splitting, in which our explanatory variable space is split in two in the way that most reduces error. This process is repeated, where each subsequent split can only happen on previously split regions. The tree is allowed to grow until the number of observations in each region reaches a threshold. It is then "pruned" back based on which splits decrease the overall error enough to justify the increased complexity of the model. That determination is made by the complexity parameter. The resulting model predicts the same mean for any data points that fall within a given region after all splits have been made. In equation form, our model is

$$\mu_t = \text{Mean}(\{y(\boldsymbol{x}_i : \boldsymbol{x}_i \in \mathcal{R}_t)\}).$$

In the equation above μ_t represents the prediction for all points falling within the region $\mathcal{R}_t$. This prediction is calculated as the mean of all observations $y(\boldsymbol{x}_i)$ which fall into the region $\mathcal{R}_t$ based on their explanatory variable values, $\boldsymbol{x}_i$.

Benefits of the tree model include its simplicity and its ability to deal with many explanatory variables. It can also find interactions, non-linear relationships, and deal with outliers more easily than many other models. Drawbacks are that the tree gives high variance estimates, does not give information on any variables not included in the tree, and it is difficult to reflect our uncertainty in tree estimates.

In light of these drawbacks we elected to use an ensemble method, the boosting algorithm. The boosting algorithm takes individual trees fitted on the residuals to that point and scales their results before adding to the overall model. Trees in a boosted model are typically more restricted in their growth than regular decision trees because each tree is fitted on the residuals of previous trees and only contributes a small fraction to the final estimate. The idea is that between all the trees, we will find all the relationships in the data and eventually our residuals will be minimized. The boosted algorithm is summarized below.

For each iteration we fit a tree $\hat{f}^b$ to the specified depth. We then update our overall model $\hat{f}$ with the scaled version of the current tree $\hat{f}(x) = \hat{f}(x) + \lambda \hat{f}^b(x)$ where λ is the scaling constant. The residuals are then updated, $r_i = r_i - \lambda \hat{f}^b(x)$ in preparation for the next tree to be fit. Once all trees have been fit, the final model estimates are the sums of the scaled estimates from each individual tree. Benefits of this model include the same benefits as regression trees, but with lower variance estimates. Drawbacks include a propensity to overfit and less interpretable output then you would get from a linear regression or a decision tree.

Model Justification and Performance Evaluation

As discussed previously, our dataset included possible explanatory variable dependencies as well as similar numbers of observations and predictor variables leading to overfitting and imprecise model estimates. We used a generalized boosted model because it repeatedly fits decision trees to create more accurate results. These iterations created by the model allow explanatory variables to be measured against multiple observations of the outcome variable, thus reducing overfitting. Boosted models can also reduce variance and bias that may have increased through variable dependecies. In order to measure the reliability of our boosted model, we performed a leave-one-out cross-validation. Our cross-validation showed that our model produced an RPMSE of 0.297 and a bias of 0.008. This shows that our model resolved previous issues and was able to fit the data with little error and reliable results.

Results

Our analysis hoped to address which variables were most important in determining river flow in the Rocky Mountains, determine how well those variables explained the observed flow, and estimate the predictive strength of those variables for data outside our sample.

To address the first goal we calculated variable importance values for each variable in our data set. These values are the reduction in mean squared error attributable to each variable, scaled so that all values sum to 100. The most important variables according to our relative influence calculations are shown below. The most important variables in the categories of climate, river network, flooded vegetation, and human influences were precipitation seasonality, drainage density, percent of flooded vegetation for land cover, and reservoir surface area respectively.

Table 1: Variable Importance

	Relative Influence
Precipitation Seasonality	18.914
Regularly flooded vegetation (%)	14.142
Cumulative January precipitation (mm)	12.029
Mean November precipitation (mm)	9.089

To determine how well our model explained the variation in observed river flow we calculated root mean squared error (RMSE) and an estimate of R^2. Our RMSE was calculated to be 0.17 and our R^2 was 0.96. Our RMSE is relatively low compared to the scale of our response and R^2 is high so we believe that the variables provide a good explanation of river flow. The most important variables identified above all have an intuitive connection to river flow as well, in that rainfall, drainage, flooding and blockages would all be expected to influence overall flow in a river.

To assess the predictive power of our variables we performed leave-one-out cross-validation and summarized the bias, out of sample R^2 and the root prediction mean square error (RPMSE). The bias and RPMSE are the average of each individual prediction's respective bias and RPMSE, while the out of sample R^2 is calculated as one minus the ratio of the sum of square prediction errors to the total sum of squared errors.

Those metrics are summarized in the table below.

Table 2: Prediction Performance Indicators

	Estimate
RPMSE	0.2974080
Bias	0.0078077
Out of sample R^2	0.7591299

Conclusions

Rivers play an important role in our ecosystem and are a major source of life. Through our analysis, we were able to identify human, river network, and climate variables that influence Rocky Mountain river flow so that farmers can make more informed agricultural decisions regarding irrigation and planting. We summarized our model's inferential and predictive power and found that precipitation seasonality, drainage density, percent of flooded vegetation, and reservoir surface area are the most important factors that influence river water flow. We now have a better understanding of which factors are related to our metric of interest, as well as how reliable those relationships are.

As mentioned in the model section of our report, the boosting model has over-fitting concerns, but we have addressed these through cross-validation. The boosting model is also less clear in interpretation than many other models. A possible next step could be to fit a regression or other interpretable model on the variables that we identified through our boosting model. Another possible next step would be to fit a bayesian autoregressive tree model, which is built on many similar ideas to boosting, but includes better quantifications of uncertainty in estimates.