Camilla Handley
Connie Mui
Stat 536
Dr. Heaton

Rocky Mountain River Drainage

Abstract

This analysis uses data from the Rocky Mountain Region of the United States to investigate factors that influence overall river flow. The dataset includes around 100 observations with 102 measured factors. We used a Principal Component regression model as well as an Elastic Net model to perform the analysis. The Elastic Net model performed better and indicated that global stream order and the amount of land cover that is Evergreen trees are the factors that influence river flow the most. This model explains river flow well and has a strong predictive capability. Although these variables adequately explain and predict river flow prices, there are other modeling techniques that could be used to analyze this data, like spatial analysis or partial least squares.

Introduction

Rivers play a very important role in all ecosystems because they carry water and nutrients to other areas. Every living organism needs water and rivers provide an excellent habitat for many of the earth's plants and animals. Rivers are especially important in the Rocky Mountain Region because farmers irrigate their cropland with water from nearby rivers. For this analysis we will use data from various rivers in the Rocky Mountain region.

The data contains a standardized measure of overall water flow as well as 97 different potential explanatory variables that could affect river flow. Among these variables are mean temperatures for each month, drainage density, length of longest path in watershed, precipitation levels, and land cover. This analysis is aimed to determine which factors have the biggest impact on overall river flow. In addition we will quantify how well these factors explain overall water flow and investigate their predictive capability.

Data Exploration

There are some problematic aspects of this data that need to be considered. When there is a very high number of covariates (close to or greater than the number of observations), various problems can arise due to the curse of dimensionality. Collinearity can make it difficult to estimate the effects and significance of the explanatory variables accurately because it will inflate the variance of the standard errors. In addition, overfitting can result in poor mean squared error (MSE) as well as false positives (i.e. associating variables that are not actually related). To work around this, we will come up with a low-dimension representation of the data to model. Methods will be discussed in the next section.

Another problematic aspect of this data are two variables, *meanPercentDC_VeryPoor* and *meanPercentDC_Well* (both percentages of land cover of a certain drain class) that have zero variance. This means that the same percentage was recorded for all 102 observations. This information does not help us in this analysis, so we will remove these variables from the data before modeling. There is also the potential for some spatial correlation between observations, but for the purpose of this analysis, we will ignore that and remove the Latitude and Longitude measurements from the data as well. We recognize that there could be value in trying a spatial analysis with this data.

Based solely on correlation, it appears that *bio15,* which measures precipitation seasonality, could be a factor that is strongly related to waterflow. It had the highest correlation out of any of the other covariates, at 0.606. Also, below is a correlation matrix with just a few of the covariates. Due to the large number of covariates, we will not plot all of them. Some possible collinearity is apparent, as well as some possible relationships between *Metric* (our response), and other covariates.

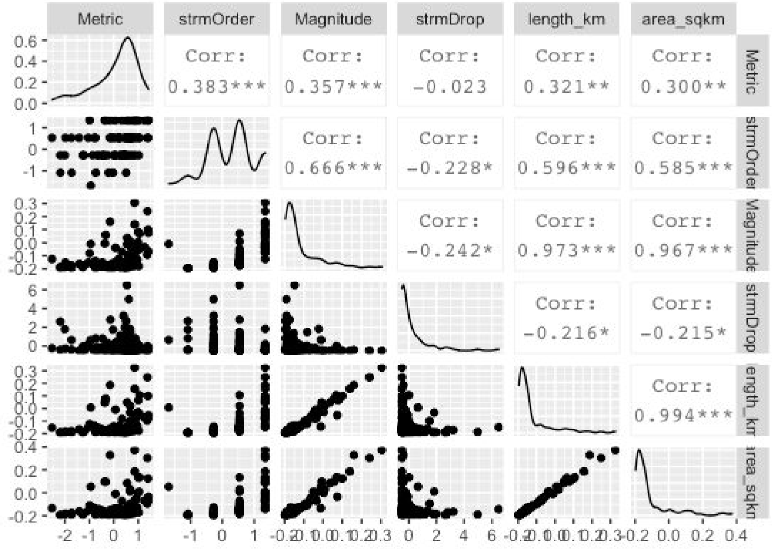

Figure 1: Correlation Matrix

Analysis Method

We chose two different models to analyze this data and compared their MSE to see which performed better. The first approach was Principal Component Analysis. This method finds the best fit line to the data that is orthogonal to the lines previous to that one. In other words, it performs a change of basis and reduces the number of components (thus dealing with the high dimension problem). Fitting a Principal Component Regression model resulted in an MSE of 0.34. The second method that we tried was an Elastic Net. This model resulted in an MSE of 0.225, so we will continue with this model for the rest of the analysis.

An Elastic Net is a special case of the classic linear model, defined below.

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}$$

Where y_i is the i^{th} observation of water flow, the x's are the observed covariates for the i^{th} observation, and β_0 - β_p are the respective coefficients. Elastic Nets use shrinkage and regularization to minimize the squared residuals as well as the coefficient size. So $\widehat{\beta}$ is calculated by minimizing the following:

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

where λ is a shrinkage parameter and the size of β is defined as the following:

$$Size(\beta_p) = \alpha|\beta_p| + (1-\alpha)\beta_p^2,\ \alpha \in [0, 1]$$

This process is also known as penalized least squares, which restricts the flexibility of the model, thus helping us with the curse of dimensionality. The shrinkage penalty increases as the number of covariates increases. In turn, we sacrifice some bias to reduce variance. The Elastic Net is a balance between LASSO and Ridge regression, making it a good choice for when some covariates are correlated, as we have in our case. Elastic Net only assumes linearity in the data.

Justification & Evaluation

The only assumption for Elastic Net models is linearity. There are no assumptions for independence, equal variance and normality. Because of the high dimensions of the data, we chose not to visually show linearity. We will proceed assuming linearity to avoid overfitting and overreacting in our model. Cross validation was used to find the best fit model based on root mean squared error (RMSE), R-squared and mean absolute error (MAE), which resulted in α = .434 and λ = .054. With these parameters, we used a bootstrap sampling technique with 1000 simulations to form 95% confidence intervals for each coefficient. The intervals that do not contain zero are the only ones that we consider significant. These can be seen in Table 1. The model produced an R-squared of 0.809, which means that 80% of the variance in water flow is explained by the selected covariates. Furthermore, the model seems to be doing an adequate job with prediction. In a leave-one-out cross validation study we found MSE = 0.225. This is very low, especially considering that *Metric* (the measure of water flow) has a standard deviation of 0.88, meaning that we are predicting well within one standard deviation. This CV study also resulted in an average bias of 0.006, which is also very low, indicating strong predictive capability.

Results

Variables	Category	Estimates	95% Confidence Intervals
intercept	*NA*	45.788	[1.433, 90.144]
gord	Network	0.284	[0.054, 0.515]
cls1	Land cover	0.126	[0.002, 0.251]

Table 1: Estimates of the Model Parameters

The table above shows the significant factors with estimates and uncertainty bounds. *Gord* represents the global stream order (predicted relationship with area) and *cls1* represents the percent of land cover that is Evergreen needle trees. These estimates can be used to predict water flow. For example, we are 95% confident that the true estimate for the effect of *gord* is between 0.054 and 0.515. *Gord* has the highest impact on water flow; as *gord* increases, so does water flow, holding everything else constant. *Cls1* has a similar effect; as the percent of land cover that is Evergreen trees increases, so does water flow, holding everything else constant. We acknowledge that there are other factors that may have a very small effect on river flow and conclude that the two factors specified above have the biggest impact on river flow. With a high R-squared, the factors were able to explain the overall water flow fairly well. Furthermore, a low MSE (.225) and a low bias (0.006) also indicate the prediction is fairly accurate.

Conclusions

Our Elastic Net model was able to answer our original research questions while dealing with the large number of covariates. The variables that we found most important in explaining river flow were *gord* and *cls1*. The model developed for this analysis was also shown to be useful in prediction. Knowing which factors affect overall water flow could help communities and farmers in this region be able to increase water flow.

This Elastic Net model was a good choice in that it has no limitation on the number of selected variables. Also, it encourages grouping effect in the presence of highly correlated covariates. However, this model suffers double shrinkage because the penalty may be higher than in Ridge or LASSO because it is a combination of the two. In the future, we want to consider spatial correlation. Finally, collecting more data is always a worthwhile consideration. This would allow us to estimate parameters with greater precision.

Teamwork

We both worked on coding the methods and compared our results and then split up the written report evenly (pointwise). We both helped edit each other's sections.