Predicting cases of dengue feaver based on evnivornemtnal data using a random forest model.

Madison Hobbs and Jenn Havens

Introduction:

Cases of dengue fever are related to the current and past climate. Environmental data can be used for predicting patterns in rates of dengue fever. Patterns of cases known in advance can used as an early warning system to help local authorities prepare for unusually high number of cases as well as informing which areas may need the most outside assistance.

The Model

We created two (2) predictive models, one for each of the cities, San Juan, Puerto Rico and Iquitos, Peru. Our model is a random forest algorithm which was trained on data from each city individually.

Missing Values:

We impute missing values using imputation via bagging (from the caret package). According to their documentation: " Imputation via bagging fits a bagged tree model for each predictor (as a function of all the others). This method is simple, accurate and accepts missing values, but it has much higher computational cost. "

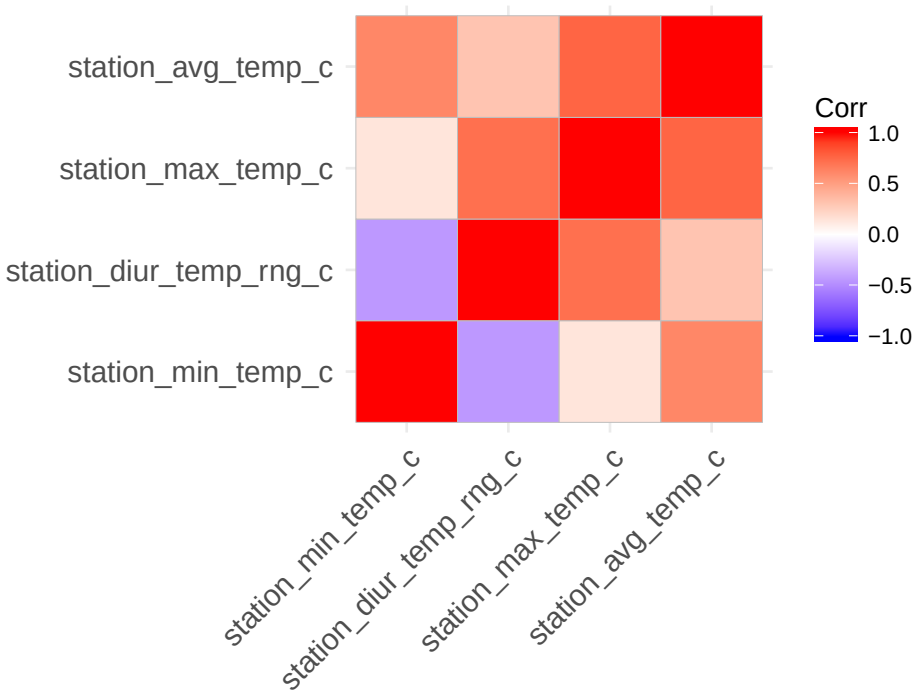
Variable Selection

Temperature

We think that the station-measured temperature would be preferable, because these are not measured from satellite or estimated from a model.

We should also think about whether to choose minimum temperature, maximum temperature, average temperature, or diurnal temperature range. We suspect that these variables are redundant.

In fact, all pairwise correlations between the four measures of temperature are significant. We can see these strong correlations represented in the correlation plot below:



Because of this correlation and because past studies have found that mean temperature was significantly associated with dengue rates, but maximum temperature and minimum were not always significant, we want to use just mean temperature for all of the correlated temperature variables.