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 numbers of cases as well as informing which areas may need the most outside assistance. DrivinData is hosting a competition: DengAI: Predicting Disease Spread, to predict cases of dengue fever in San Juan, Puerto Rico and Iquitos,Peru.

The Model

We created two (2) predictive models, one for each of the cities of interest: San Juan and Iquitos. Our model is a random forest algorithm which was trained on data from each city individually. A random forest algorithm builds up many decision trees which use explanatory variables to separate the possible responses in a response variable based on training data and then uses the average of the predicted results from these trees to predict new data. The response that we are interested in is the number of dengue fever cases. The explanatory variables in this model are different measures of environmental conditions including temperature, humidity, precipitation, and Normalized Difference Vegetation Index (ndvi). We also considered time of year as a possible variable. The environmental training data was collected from the US National Oceanic and Atmospheric Administration (NOAA) and the health data (ie cases of dengue fever reported) from the US Centers for Disease Control and Prevention (CDC).

Missing Values:

First, we ask how many values in our data are NA?

```
## [1] 0.015
```

Only 1.5% of our data is missing. Because there are relatively few missing values, and because having missing weeks will complicate our time series analyses, we decide to impute missing values. We use 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. " Luckily, it doesn't take too long for bagging to impute the relatively few missing values on our data.

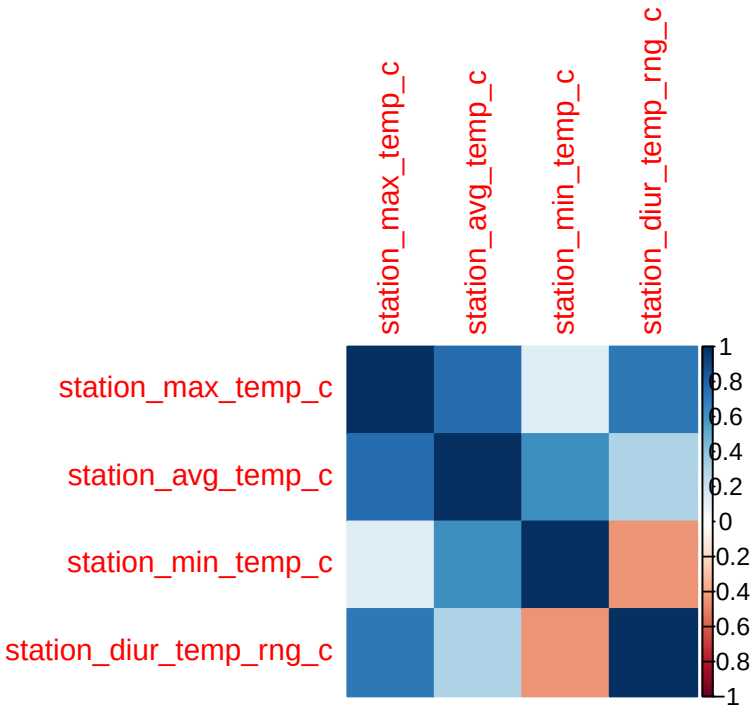
Variable Selection

Temperature

We think that the station-measured temperature would be preferable, because these are not measured from satellite or estimated from a model. The NOAA Dengue Forecasting project also notes in their reference guide, "Environmental data sources for the Dengue Project," that, "Ground observations are generally an optimal representation of actual local conditions."

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 a past study have found that mean temperature was significantly associated with dengue rates, but maximum temperature and minimum were not always significant, we will proceed with mean temperature as the only temperature variable.

Precipitation

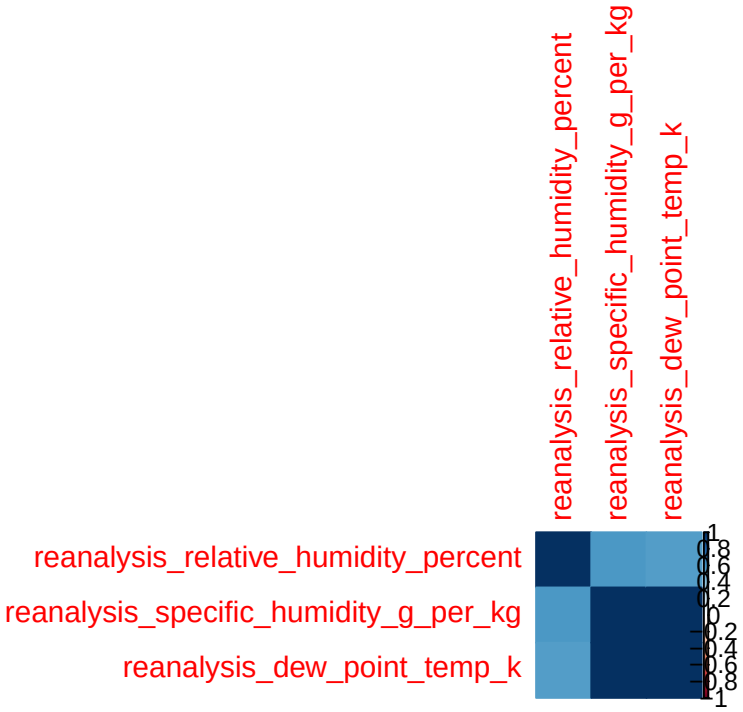
Variable descriptions are avalible.

There are three measures of precipitation. One is station_precip_mm which is the total daily precipitation as measured by NOAA's GHCN weather stations. Another is precipitation_amt_mm which represents total precipitation as measured by PERSIANN satellites. The third and forth, reanalysis_sat_precip_amt_mm and reanalysis_precip_amt_kg_per_m2 are both generated by NOAA's NCEP Climate Forecast System Reanalysis.

We should choose one of these precipitation measures for the model, since these four precipitation measures all measure approximately the same thing. According to the NOAA Dengue Forecasting recommendations, "remotely sensed observations are generally an excellent observation of precipitation and vegetation conditions for a location." With the multiple sources for total precipitation, we decide to use only the satellite measured total precipitation for each city to build our model.

Humidity and Dewpoint

Specific humidity, relative humidity, and dew point are all provided and are all measured using NOAA's NCEP Climate Forecast System Reanalysis. All three measures are significantly correlated, as seen below.



Since these measure roughly the same information, we opt to include only one in our model. Relative humidity is the measure most often used in literature we have read. It is also an easy-to-find measure, making our model more user-friendly. Therefore, we use only relative humidity in the model.

Normalized Difference Vegetation Index: NDVI

NDVI is an indicator of vegetation by measuring the amount of live green plant material in an area as seen by satellite. Past studies have found that measures of vegetative indices are correlated with rates of dengue fever. It has been proposed that vegetation provides an environment for mosquitoes, vectors of dengue fever to lay eggs. Vegetation can also effect the temperature, precipitation, and humidity of micro-environments. We have used used the ndvi measured at the four points closest to the city's central point, one in each direction.

Creating Lagged Data

We separate San Juan and Iquitos data to produce two models, one to predict weekly dengue fever cases in San Juan and the other in Iquitos. This is because weather and vegetation will behave differently in relation to time between both locations, because these are locations separated by distance, climate, and ecosystem.

At the same time, we create two lag variables for each site: temperature lag and precipitation lag. The lagged temperature and precipitation variables record, respectively, what the temperature or precipitation was 12 weeks prior for the observation at hand.

Time lagged environmental variables have been shown to be a signifigant predictors. Specifically we consider temperature, precipitation, and relative humidity.

Model Evaluation

Michael J. Kane et. al compared a time series Random Forest model and an ARIMA model to predict avian influenza H5N1 cases, detailed in their paper. They found that a Random Forest time series model with time lag variables out-performed the prospective ARIMA model in predicting H5N1 cases per week in Egypt.

To assess their model, Kane et. al built a Random Forest model on 30 weeks, then used that model to predict the next week. The simulation steps forward week by iteratively adding the next week of data, building a new model, and predicting the number of cases in the following week.

Inspired by their work, we was used for prediction of avian influenza H5N1 outbreaks. The model was assessed by taking data from the previous 30 weeks to predict for the next week. The simulation steps forward iteratively adding new data for each week, the new model using the updated data to predict the next week, and comparing to the true number of cases.

Moving in steps of 30 weeks, we train the model on 30 weeks, then predict the next week.

We'll assess our model's performance by computing the error (root mean squared error) between the model's predicted number of cases and the actual number of cases for each week. The Driven Data team uses mean absolute error to assess models, so we will, too.

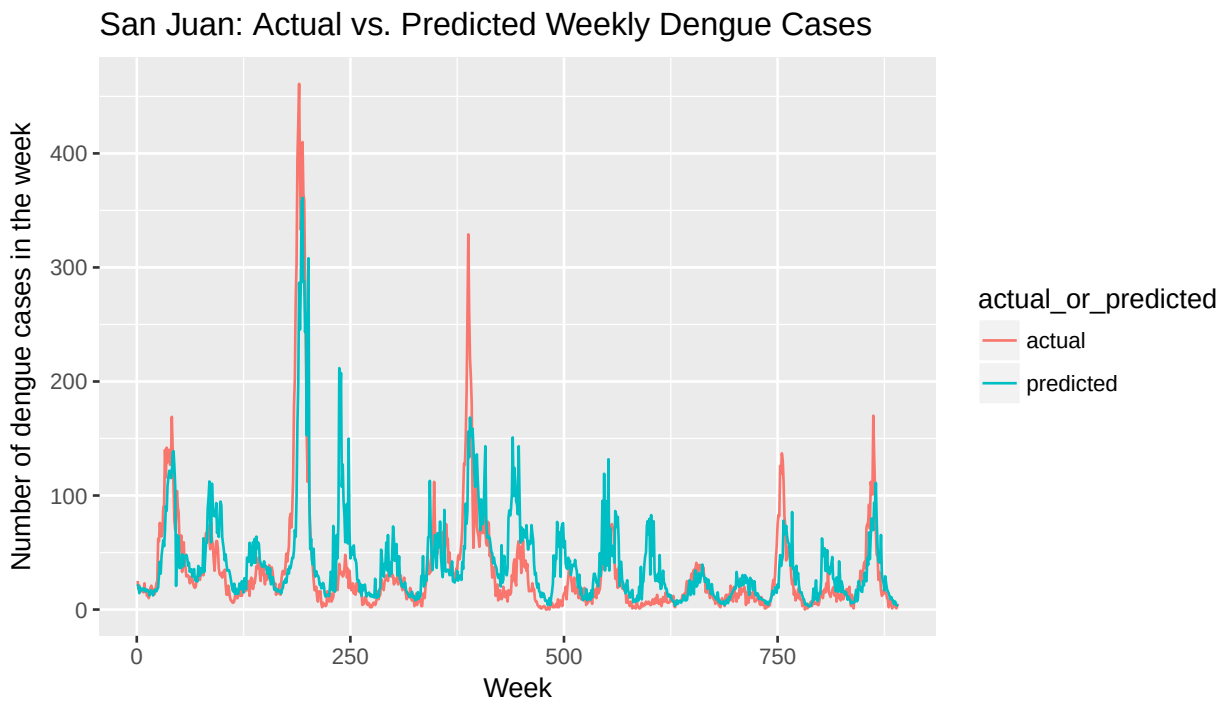
San Juan

Testing with current environemntal data, lagged environmental data, and date information

The normalized Root-Mean Square Error of the model with current environmental data, lagged environmental data, and date information on the San Juan training data is:

```
## [1] 67.2
```

This is a relatively high error rate. We can visualize how the predicted number of cases matches with the actual number of cases.



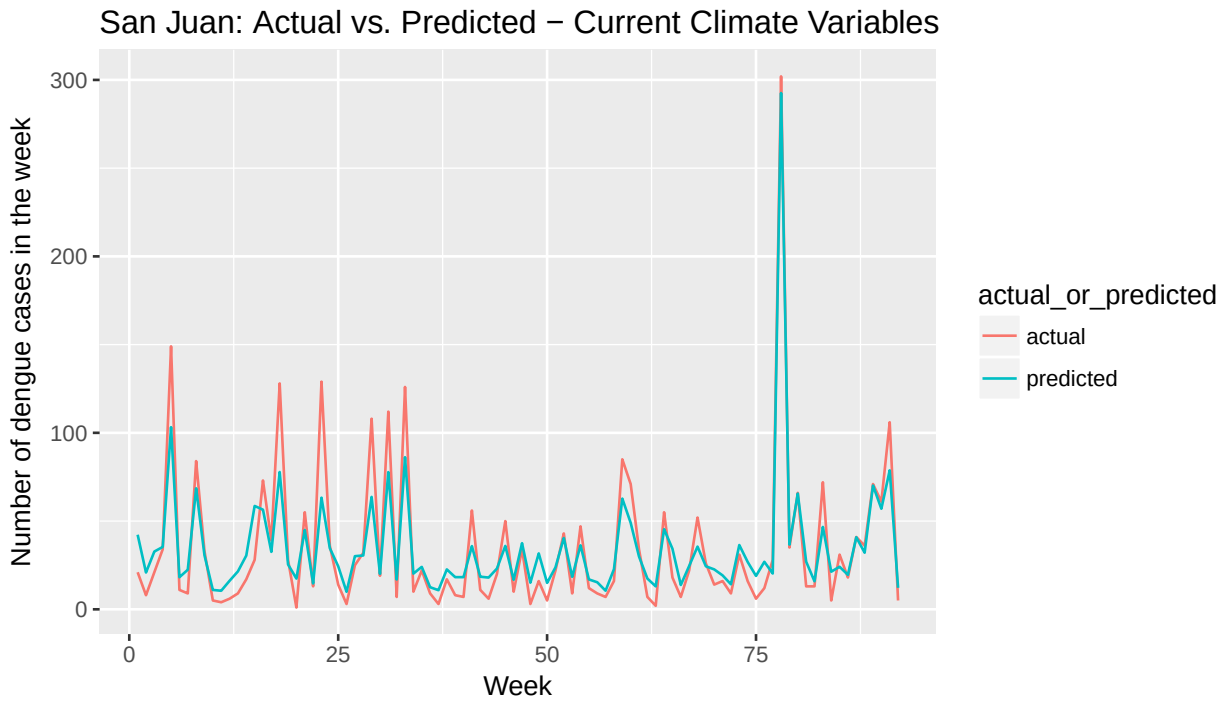
Based on this we can see that the model has good predictive ability, on the included data. In some outbreaks the absolute count predicted by the model was less than the true number of cases. However, the model was able to predict timing of outbreaks well, and relative size of the outbreaks.

With current climate varibles only

The normalized Root-Mean Square Error of the model for San Juan with only current environmental data is:

```
## [1] 24.1
```

The normalized prediction error rate for this model is better than the model better than the model with current environmental data, lagged environmental data, and date information.



Comparing predicted to actual cases in a subset of the weeks, with the model using only current climate conditions, we see similar patterns in the prediction to the model which used current climate conditions, past weather conditions, and the time of year.

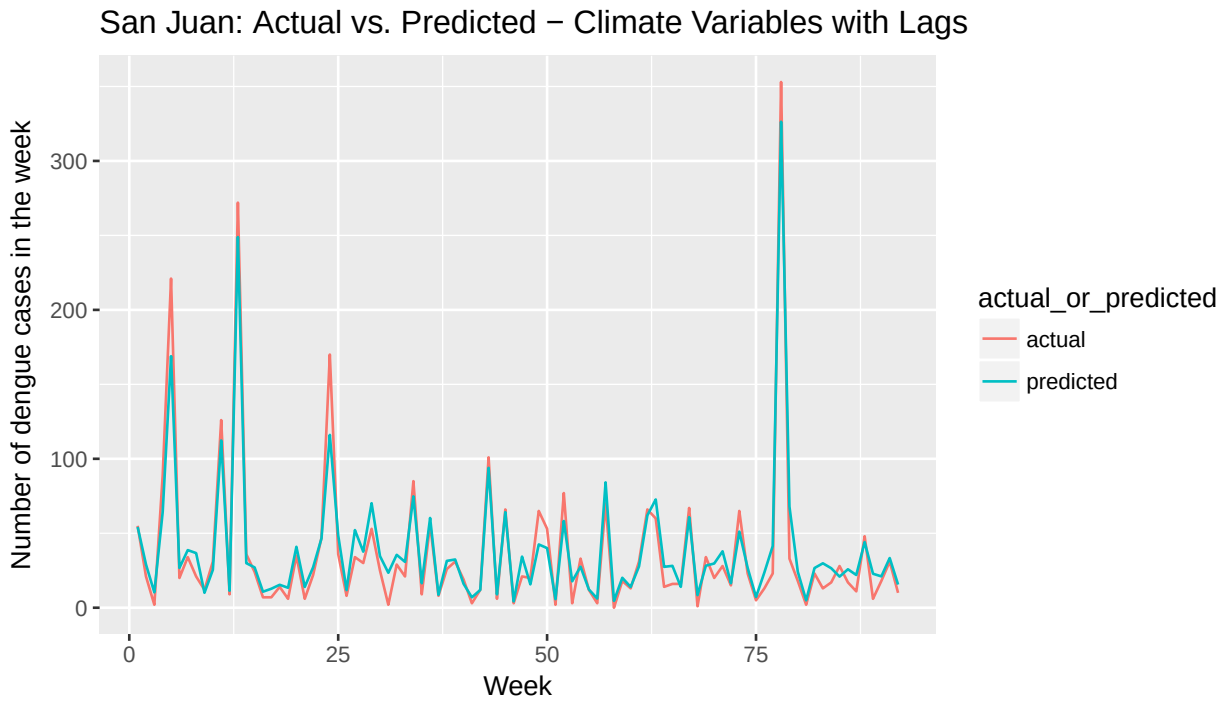
With Current Climate and Lagged Weather

This models includes current ndvi, current temperature, humidity, and precipitation, and lagged temperature, humidity, and precipitation

The normalized Root-Mean Square Error of the model for San Juan with current environmental data and lagged weather data:

```
## [1] 24.1
```

The normalized error rate is better for the model with current environmental data and lagged weather data than the other models.



Comparing predicted to actual cases in a subset of the weeks, with the model using current climate conditions and lagged weather conditions, we see similar patterns in the prediction to the model which used current climate conditions, past weather conditions, and the time of year. However this model has an improved RMSE, and the variables used all have an association with causality for outbreaks, so this is the model that we can use for predicting cases in San Juan.

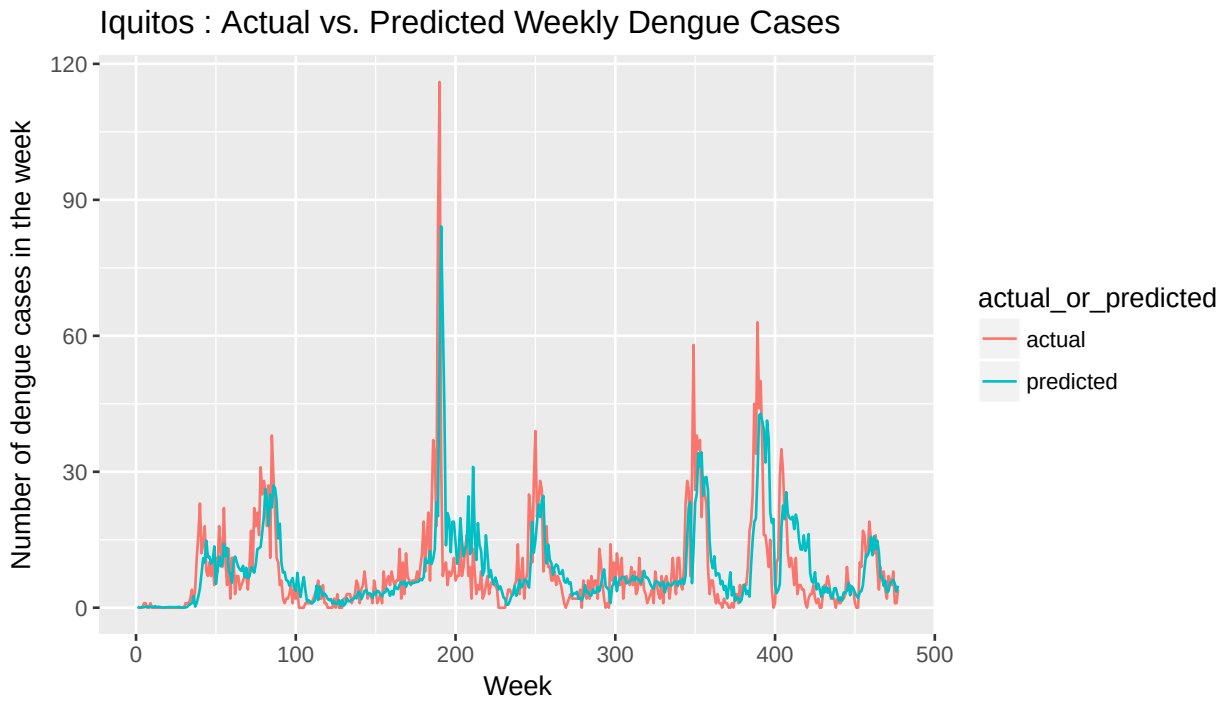
Iquitos

Testing with current environemntal data, lagged environemntal data and date information

The normalized Root-Mean Square Error of the model with current environmental data, lagged environmental data, and date information for modeling cases in Iquitos:

```
## [1] 82.7
```

This is a relatively high error rate. We can visualize how the predicted number of cases matches with the actual number of cases.



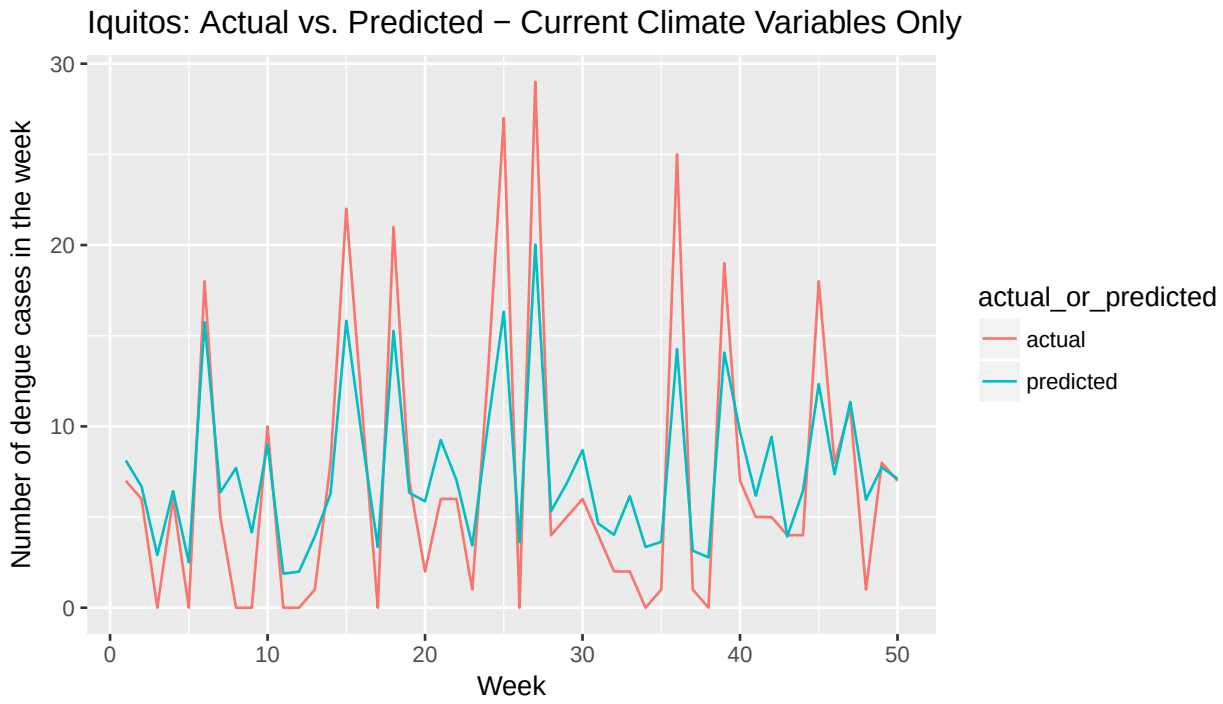
Based on this we can see that the model has good predictive ability, on the included data. In some outbreaks the absolute count predicted by the model was less than the true number of cases. However, the model was able to predict timing of outbreaks well, and relative size of the outbreaks.

With environmental varibles only

The normalized Root-Mean Square Error of the model with current environmental data only for modeling Iquitos:

```
## [1] 49.2
```

The normalized prediction error rate for this model is better than the model better than the model with current environmental data, lagged environmental data, and date information.



Comparing predicted to actual cases in a subset of the weeks, with the model using only current climate conditions, we see similar patterns in the prediction to the model which used current climate conditions, past weather conditions, and the time of year.

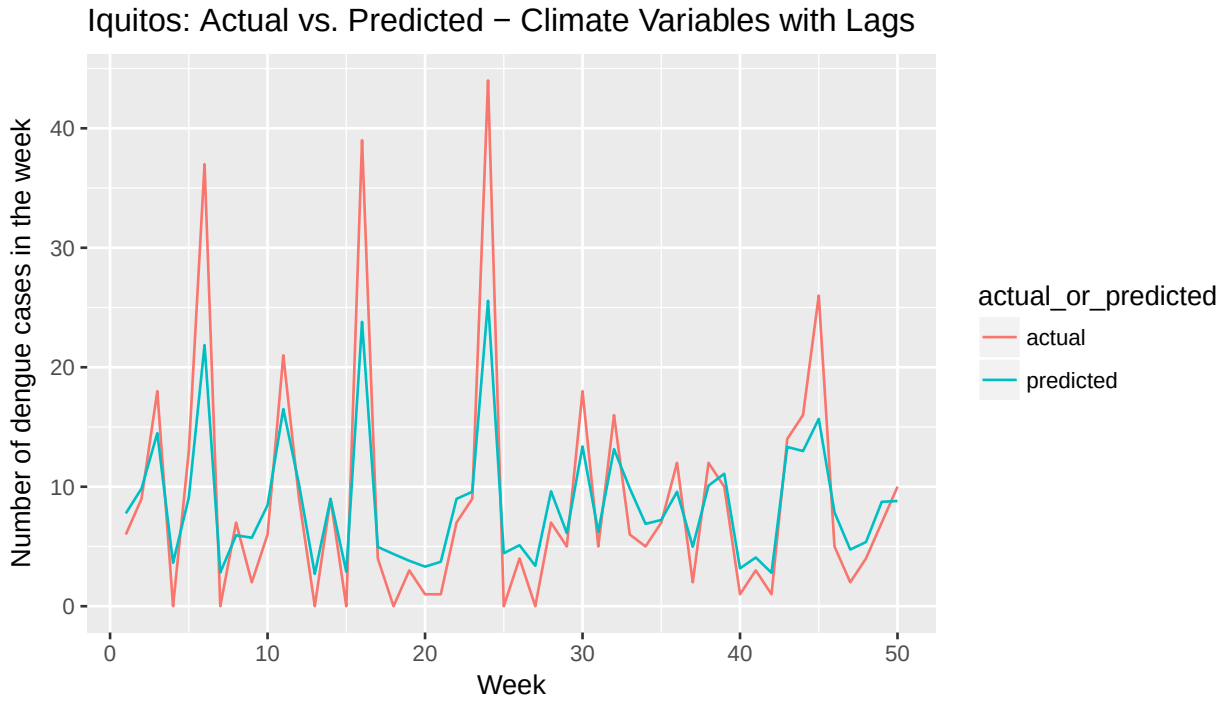
With Lagged Temperature, Humidity, and Precipitation

The normalized Root-Mean Square Error of the model for Iquitos with current environmental data and lagged weather data:

```
## [1] 49.2
```

The normalized error rate is better for the model with current environmental data and lagged weather data than the other models.

Figure 1:



Comparing predicted to actual cases in a subset of the weeks, with the model using current climate conditions and lagged weather conditions, we see similar patterns in the prediction to the model which used current climate conditions, past weather conditions, and the time of year. However this model has an improved RMSE, and the variables used all have an association with causality for outbreaks, so this is the model that we can use for predicting cases in Iquitos.

App and user prediction

We created a user interface to get predictions based on the models developed above. We learned to build a shiny app, so that the user could have a graphical interface. These models are developed specifically for the two locations for which we got the training data from, San Juan and Iquitos. The user has the option to enter current and lagged environmental data and the current NDVI. They also have the option to enter a date and the weather data at that date and lagged data from will be collected from Weather Underground. This data is collected using an API, that we learned to access in the course of this project. The NDVI used in this prediction will be the average of the training data.

Conclution

The best predictive models for cases of dengue fever in San Juan, Puerto Rico and Iquitos,Peru were found using current climate data (weather and vegetative) and weather data from previous weeks (lagged data) as explanatory variables in a random forest model. We were able to use theses models as the basis of a tool which allows someone to predict cases of dengue fever based on the environment. The environmental conditions are thought to contribute to patterns in dengue fever outbreaks because environmental conditions effect mating and survival of mosquitoes, which are vectors for dengue fever. We have shown that just a few metrics of environmental conditions is sufficient to give a rough indication of the expected out breaks in San Juan, Puerto Rico and Iquitos,Peru. By using different training data it would be possible to develop models for any area of interest. This could be useful in prioritizing which areas receive assistance in possible large scale outbreaks. This model is limited by considering only the past environmental situation. It is has been proposed that in the future current patterns of outbreaks will be disrupted, as climate change effects regular weather patterns and urbanization disturbs mosquito habitat (standing water where mosquitoes can lay eggs). The burden of dengue fever has been estimated to have increased 30-fold in 50 years, and the increase is expected to continue if there is no intervention.