Predicting cases of dengue fever based on environmental 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. Understanding how environmental factors predict dengue fever rates can serve as an early warning system to help local authorities prepare for unusually high numbers of cases and inform them about which areas may need the most outside assistance. DrivenData is hosting a competition: DengAI: Predicting Disease Spread, to predict cases of dengue fever in San Juan, Puerto Rico and Iquitos, Peru.The data comes from the NOAA Dengue Forecasting project.

Model Approach

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

A random forest algorithm builds up many bagged and decorrelated decision trees which use explanatory variables to group the possible values of a response variable based on training data. To predict an observation with unknown response, we send that observation (and its explanatory variables) though all of those decision trees, get an answer from each one, then take the average of those answers to be our prediction.

The response that we are interested in is the number of dengue fever cases in a week. 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 as a possible variable, constructing a time-series random forest. The environmental training data was collected from the US National Oceanic and Atmospheric Administration (NOAA) and the health data (cases of dengue fever reported) from the US Centers for Disease Control and Prevention (CDC). For more information, please see "variable_descriptions.pdf" on our github.

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." Fortunately, it doesn't take too long for bagging to impute the relatively few missing values on our data.

Variable Selection

Some variables, like temperature and precipitation, are provided to us from multiple sources in the Driven Data dengue fever prediction data set. Some data comes from NOAA GHCN station measurements in San

Juan, Puerto Rico and Iquitos, Perú. See NOAA's "Environmental data sources for the Dengue Project" on our github for specific geographical coordinates of these stations.

A second data source is NOAA's CDR PERSIANN Precipitation Product which uses remote sensing (satellites) coupled with an artificial neural network to produce precipitation data for every 0.25 x 0.25 degree on Earth.

A third data source is the Climate Forecast System Reanalysis (CFSR) provides the "best estimate" for the state of the "atmosphere-ocean-land surface-sea ice system" over a given time period (according to NOAA via "Environmental data sources for the Dengue Project"). It is from this source that our humidity and dew point measurements come.

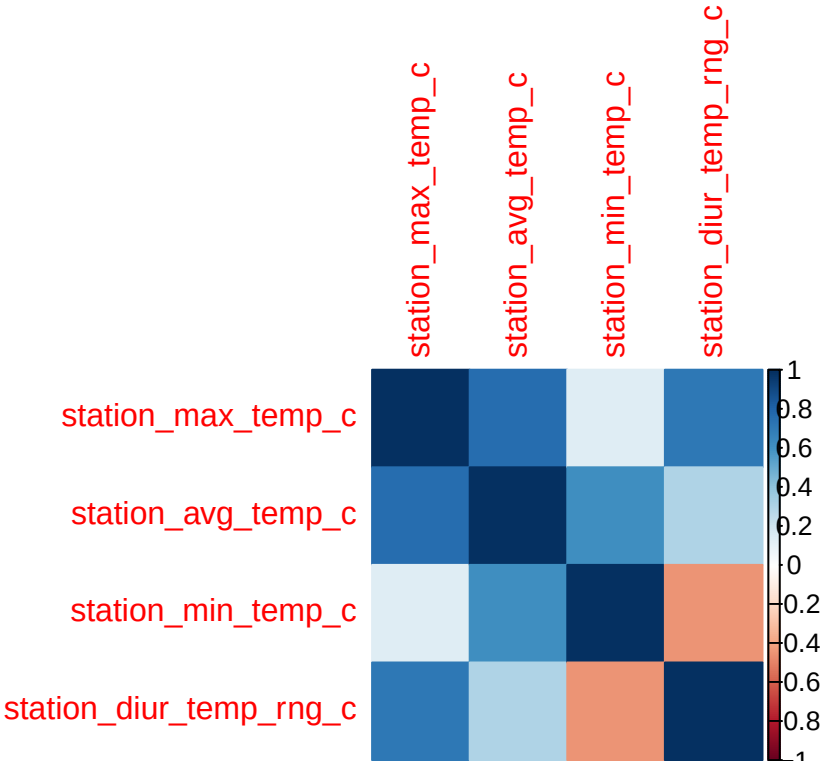
Temperature

There are two different sources for temperature: station and Climate Forecast System Reanalysis.

We think that the station-measured temperature is the preferable predictor because these are not estimated from a model, but rather recorded directly. 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."

Stations provide minimum temperature, maximum temperature, average temperature, or diurnal temperature range. We suspect that these variables are redundant and investigate their correlation below.

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

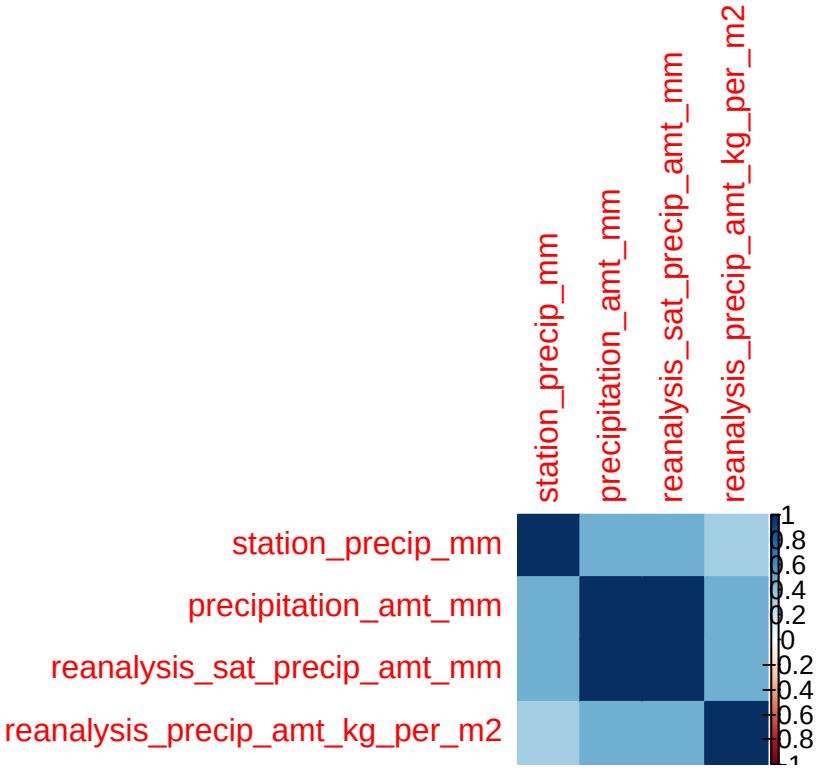


Because multiple correlated variables are redundant and ill-advised to include in a random forest model, we decide to use only station-measured average temperature. The average captures the variability of the other temperature measures. Furthermore, one past study found that mean temperature was significantly associated with dengue rates, but maximum temperature and minimum were not always significant. Therefore, we will proceed with mean temperature as the only temperature variable.

Precipitation

Precipitation is measured as total precipitation in a week, given in three different sources. One is station-measured (`station_precip_mm`) which is the total weekly precipitation as measured by NOAA's GHCN weather stations. Another is total weekly precipitation as measured by PERSIANN satellites and model (`precipitation_amt_mm`). 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.

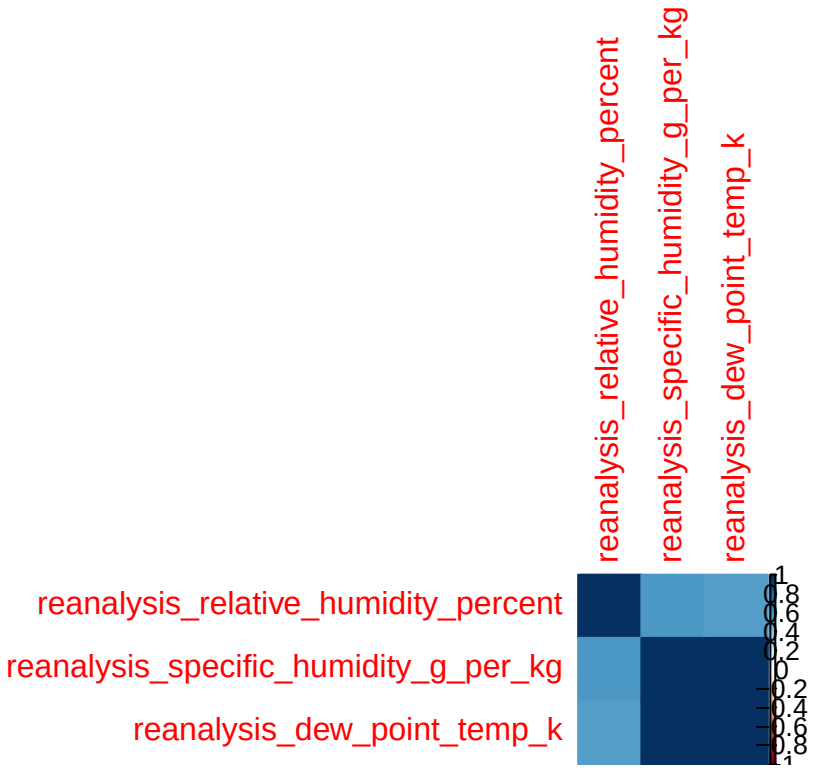
Again, we should choose one of these precipitation measures for the model, since these four precipitation measures all measure approximately the same thing, as seen below. All pairwise correlations are found to be positive and significant.



According to the NOAA Dengue Forecasting recommendations, "remotely sensed observations are generally an excellent observation of precipitation and vegetation conditions for a location." We therefore decide to use the PERSIANN satellite-measured total precipitation for each city as the only precipitation variable in 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 positively correlated, as seen below.



As before, 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, 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 affect 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, population, and ecosystem.

Time lagged variables of 2-3 months have been shown to be signifigant predictors. Specifically we consider temperature, precipitation, and relative humidity.

We therefore create three lag variables for each site: temperature lag, precipitation lag, and humidity lag. The lagged variables respectively record what the temperature, precipitation or humidity was 12 weeks prior to the observation at hand.

Model Evaluation

For each site, we construct and assess three versions of the random forest model:

1) Random Forest with Current Environmental Data, Lagged Weather Data, and Date Information
   - "current" environmental factors as predictors
   - weather time lags as predictors
   - year and week as predictors
2) Random Forest with Current Climate Variables Only
   - "current" environmental factors as predictors
   - no time lags
   - year and week not included as predictors
3) Random Forest with Current Climate Variables with Lagged Weather Data
   - "current" environmental factors as predictors
   - weather time lags as predictors
   - year and week not included as predictors

More on the Time Series Random Forest

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.

Using the same method, we constructed a time series random forest model, using time as a predictor and included the time lags created above. We use only 100 trees per forest, because run-time is otherwise a nightmare. We assessed our model as they did:

1) Start by building a model on the first 30 weeks of data
2) Use the model just built to predict the 31st week of data
3) Write down that prediction for the number of dengue cases in week 31
4) Now train a random forest on the first 31 weeks of data (actual values, no predicted values)
5) Use the model just build to predict the 32nd week of data
6) Write down that prediction for the number of dengue cases in week 32
7) Now train a random forest on the firs 32 weeks of data ...

... and so on until there are no weeks left to predict!

In the end, we get n - 30 independent predictions for dengue cases in a week (we get predictions for all but the first 30 observations in the training set). We compare these predictions to the actual number of dengue cases and calculate error.

Measuring Error

We use root mean squared error (RMSE) and Normalized RMSE (NRMSE) to compare actual and predicted values. RMSE helps us get a sense of how "off" our prediction might be. NRMSE helps us compare error

across different models and the two data sets. NRMSE is root mean squared error as a percentage of the standard deviation of the observations and thus ranges from 0 to 100 (see the nrmse function). Normalizing RMSE in this way helps us compare error across different models or data sets with different scales. We have two data sets with different scales (San Juan experiences many more dengue cases than Iquitos), so NRMSE will be useful to us.

San Juan

1. Random Forest with Current Environmental Data, Lagged Weather Data, and Date Information

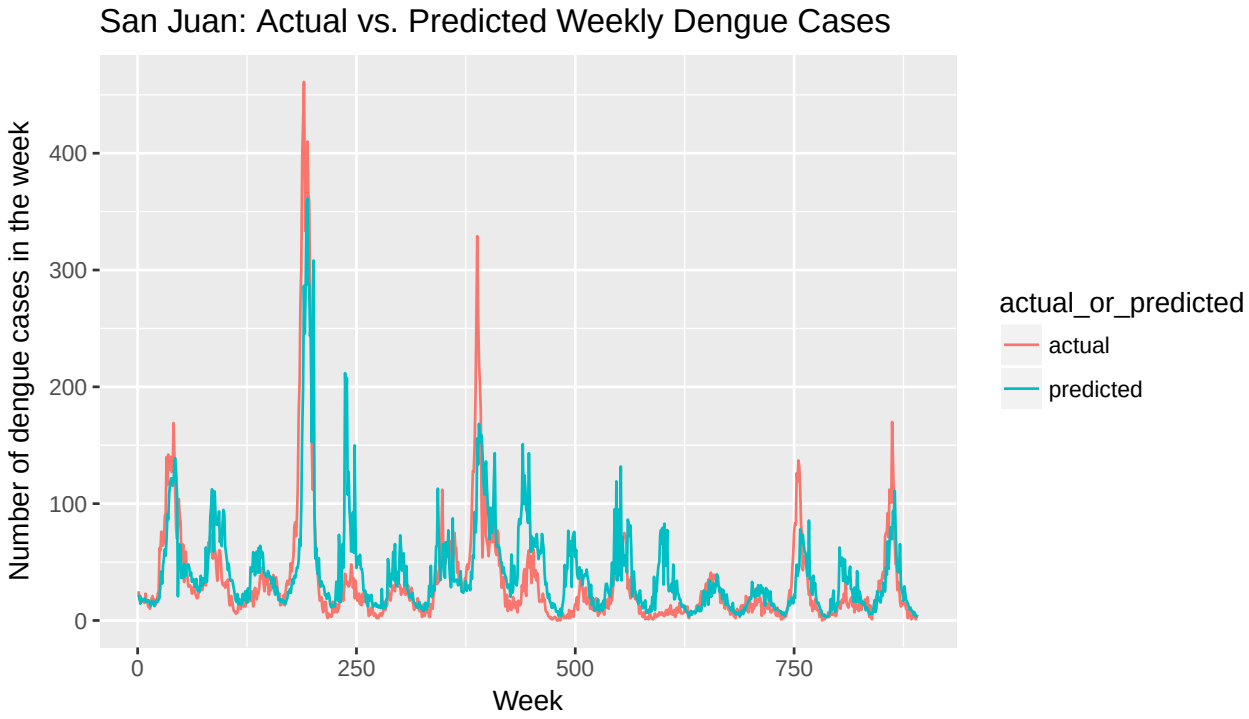
The RMSE of the model with current environmental data, lagged environmental data, and date information on the San Juan training data is:

```
## [1] 35.26671
```

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, while sometimes the model predicted outbreaks much larger than observed. However, the model was able to predict timing of outbreaks well.

2. Random Forest With Current Climate Variables Only

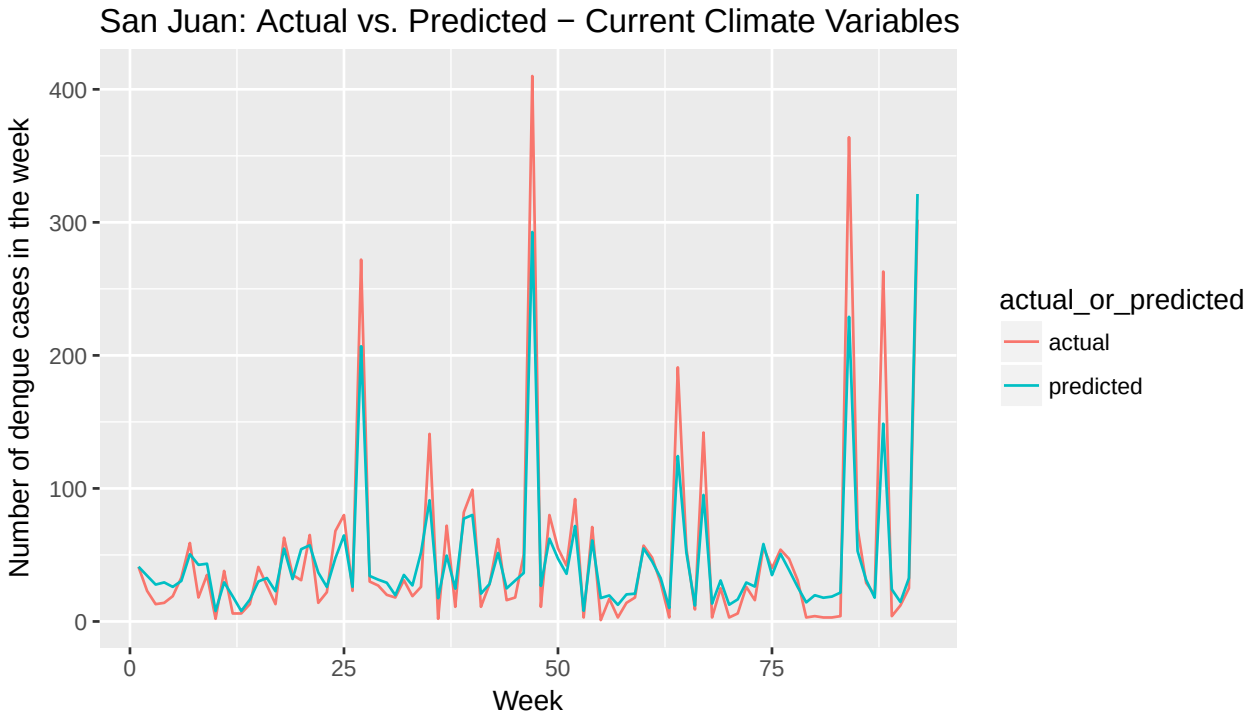
The RMSE of the model for San Juan with only current environmental data is:

```
## [1] 27.51231
```

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

```
## [1] 36.9
```

The normalized prediction error rate for this model is much better than the model better than the model with current environmental data, lagged weather 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. This model predicts relative size of outbreaks apparently better than the random forest with time variables as predictors.

3. Random Forest with Current Climate Variables and Lagged Weather Data

We suspect that we can get improved model performance by including time lagged variables.

This model includes current NDVI, current temperature, humidity, and precipitation, as well as lagged temperature, lagged humidity, and lagged precipitation.

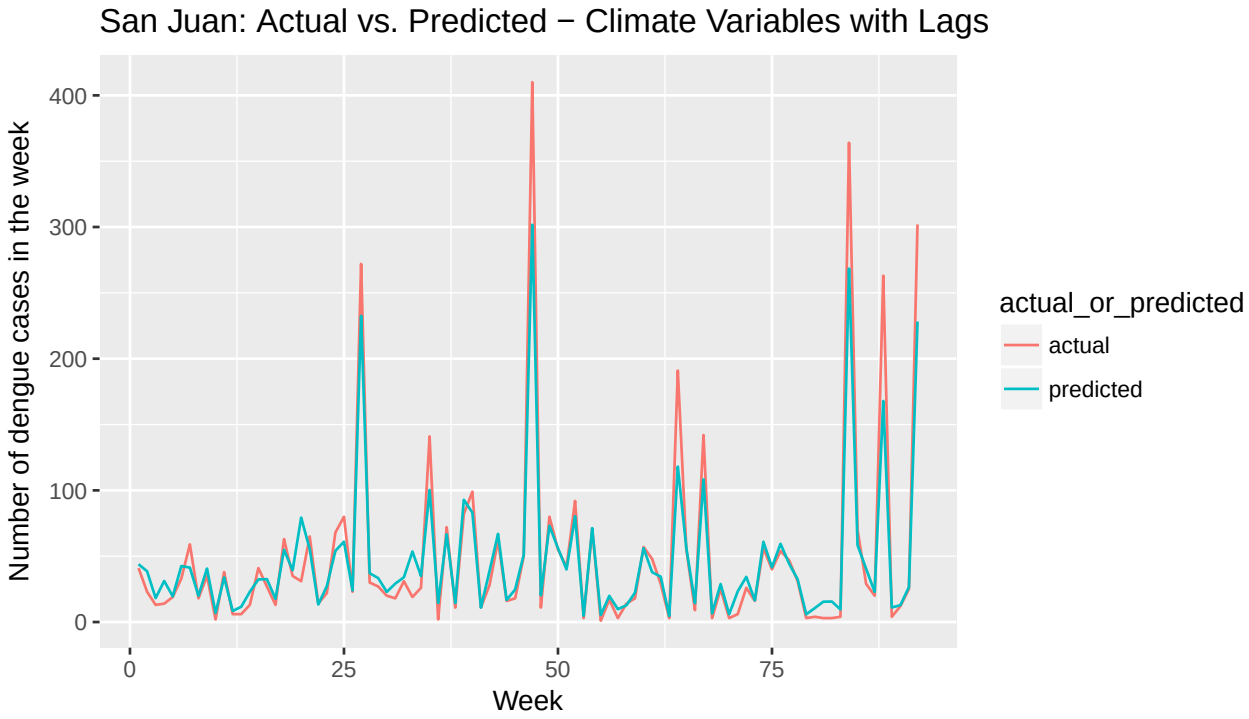
The RMSE of the model for San Juan with current environmental data and lagged weather data is:

```
## [1] 24.09376
```

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

```
## [1] 32.3
```

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 the lowest RMSE and NRMSE, and the variables used all have scientifically hypothesized causality to outbreaks, so this is the model that we can use for predicting cases in San Juan.

Iquitos

1. Random Forest with Current Environmental Data, Lagged Weather Data, and Date Information

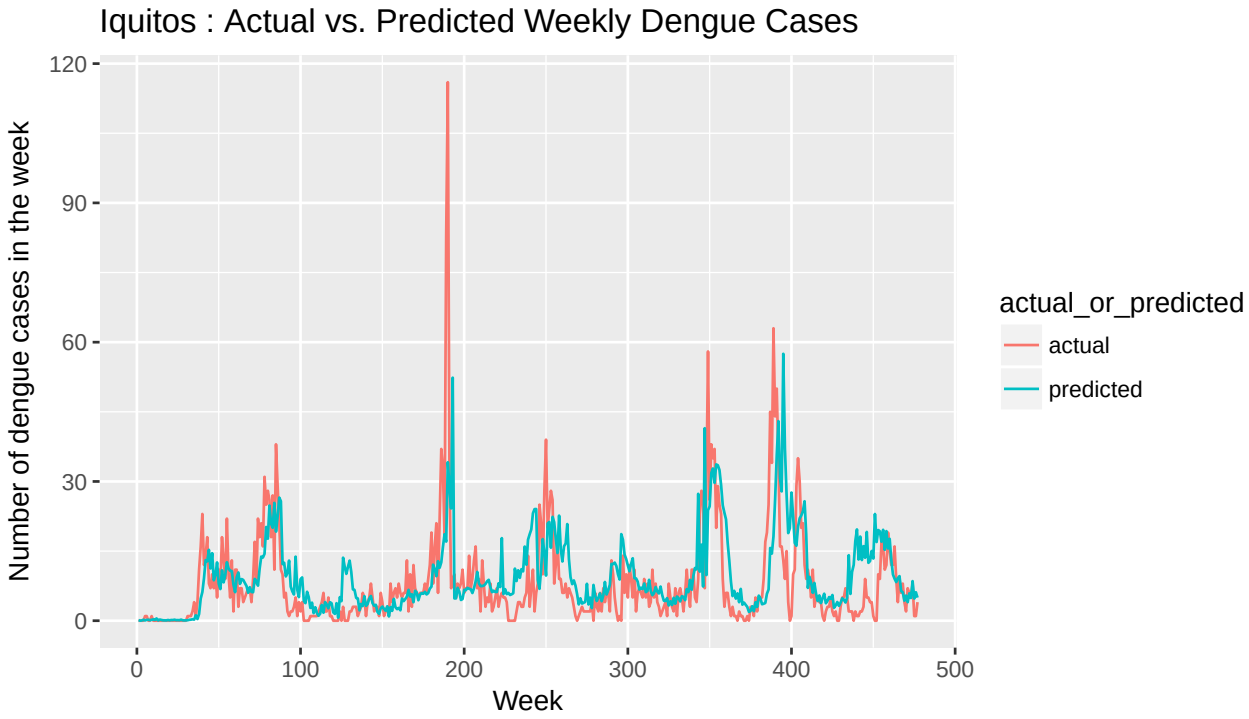
The RMSE of the model with current environmental data, lagged weather data, and date information for modeling cases in Iquitos is:

```
## [1] 9.758371
```

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

```
## [1] 88.7
```

Again, we see a relatively high error rate the random forest with date information. 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 some 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.

2. Random Forest With Environmental Variables Only

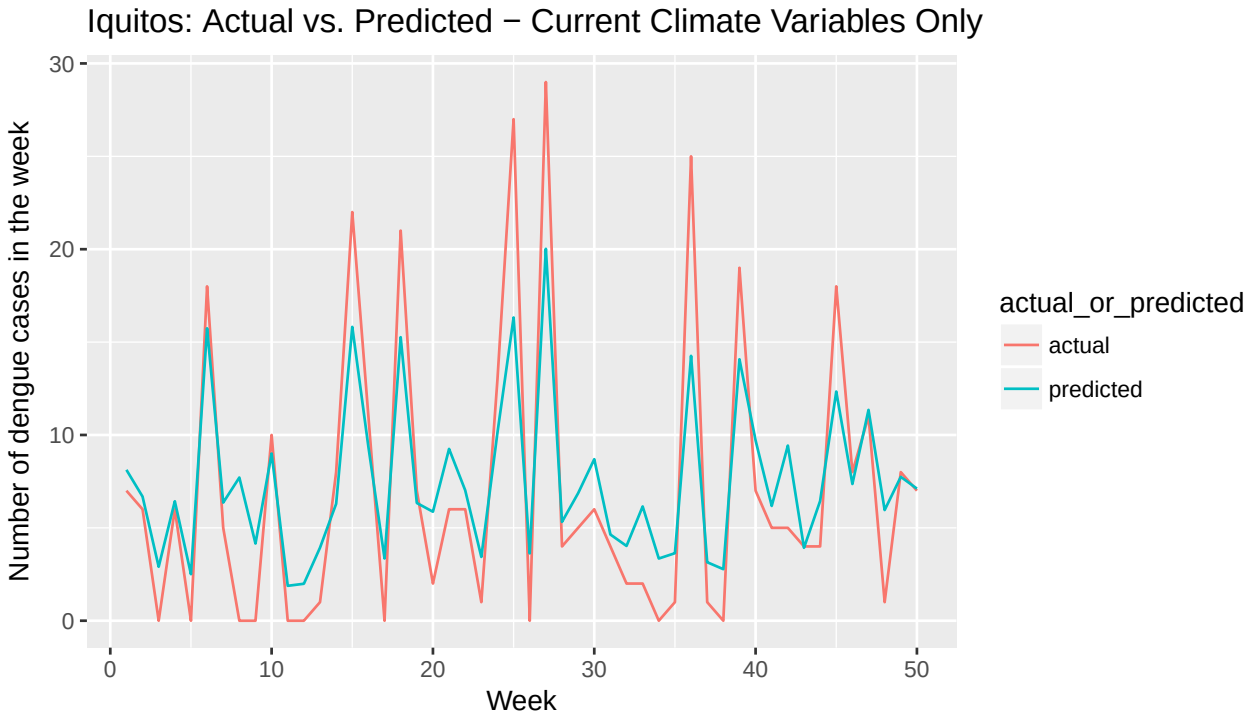
The RMSE of the model with current environmental data only for modeling Iquitos is:

```
## [1] 3.873413
```

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

```
## [1] 50.4
```

The normalized prediction error rate for this model is 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.

3. Random Forest with Current Climate Variables and Lagged Weather Data

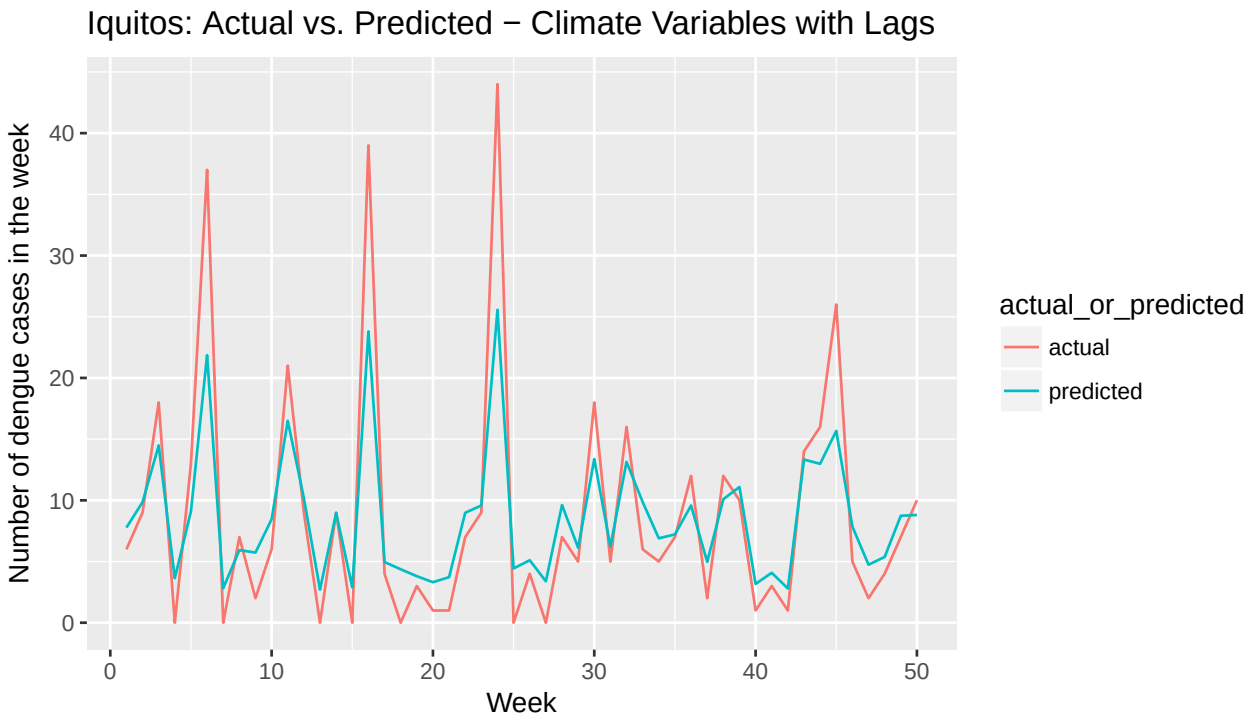
The RMSE of the model for Iquitos with current environmental data and lagged weather data is:

```
## [1] 4.910832
```

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

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

The normalized error rate is better than either of 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 yields the best RMSE and NRMSE, and the variables used all are scientifically hypothesized to have 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 an interactive 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. Alternatively, they can 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 our training data.

Figure 1:

Conclusion

The best predictive models for cases of dengue fever in San Juan, Puerto Rico and Iquitos, Perú 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 these models as the basis for 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 affect mating and survival of mosquitoes, which are vectors for dengue fever. We have shown that just a few metrics of environmental conditions are sufficient to give a rough indication of the expected outbreaks in San Juan, Puerto Rico and Iquitos, Perú. 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, and only the years 1990 to 2008 (San Juan) and 2000 to 2010 (Iquitos). It is has been proposed that in the future, 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.

Future Areas of Work

We used simple test and training to assess the two models without time variables as predictors. An extension to our work would include a more comprehensive approach such as iterative model construction and assessment (as performed on the model with time variables as predictors).

We also could have played around with the length of the time lags. 2-3 months was the suggested length according to literature we read (as mentioned above), so we chose 12 weeks. However, we could approach the length of time lag as a tuning parameter and select the time lag for each climate variable which generates the most accurate predictions.

As mentioned above, the data we used to build and assess our models represents a fraction of the available data on the web. A good follow-up would be to compile all dengue fever and climate data available, or at least the most recent data (post-2010) to build more informed models for San Juan and Iquitos. It would also be useful to build models for other areas affected by dengue fever.