Predicting Dengue Fever Outbreaks

Anuj Anand, Smit Kiri, Nitin Kumar Mittal, Nicole Reitz, Fenil Shah

December 08, 2019

Summary

The project is inspired by an online challenge on DrivenData, a website that facilitates socially impactful data science competitions [1]. The challenge is to use climate data to predict the number of weekly dengue fever cases in Iquitos, Peru and San Juan, Puerto Rico.

Dengue fever is a mosquito-borne illness that is common in more than 110 countries worldwide. With more than 60 million cases reported annually, the disease causes a painful rash, flu-like symptoms, severe muscle and joint pain, and in some cases, death [2]. Dengue is carried by the Aedes genus of mosquito, species of which are responsible for the spread of several other tropical diseases including the Zika Virus and Yellow Fever [3]. As mosquito populations are closely tied to climate conditions, it may be possible to use climate data to predict the prevalence of mosquito-borne illnesses. Identification of appropriate climate variables to model mosquito-borne illness will allow researchers and medical professionals to prepare for the global health effects of climate change [4].

The data provided by DrivenData is consolidated from multiple sources by the National Oceanic and Atmospheric Administration (NOAA) [5]. The dataset contains 1,456 rows of data; 520 for Iquitos and 936 for San Juan. The dataset contains weekly observations of reported dengue cases and climate measurements for years 2000-2010 for Iquitos and 1990-2008 for San Juan. Climate variables include, but are not limited to:

- Mean dew point temperature (K)
- Mean relative humidity (percentage)
- Mean specific humidity
- Average temperature (C)
- Diurnal (daily) temperature range (C)
- Maximum temperature (C)
- Minimum temperature (C)
- Total precipitation (mm)

We explored and compared several predictive models, including univariate time series methods and a multivariate neural network. The process revealed the strengths and weaknesses of each approach in terms of implementation complexity and predictive value. In keeping with the competition outlined on DrivenData, we used Mean Absolute Error (MAE) as a performance metric to evaluate our models.

Methods

We used both R and Python for exploratory data analysis and implemented selected models in Python using the StatsModels, Prophet, and Keras libraries.

Exploratory Data Analysis

The project began with an exploratory data analysis to identify climate factors potentially influencing the number of dengue cases reported each week. Given the temporal nature of the data, our first question was whether or not the data was seasonally distributed. As Figure 1 suggests, dengue cases do initially appear to be distributed seasonally, with most cases occurring between October to February in the city of Iquitos and August to December in San Juan across all available years of data [Figure 1]. However, closer inspection reveals that outbreak timing varies year to year in both cities, and is not quite as predictable as traditionally defined seasonality [Figure 2]. Also, some years show an irregularly high number of dengue cases, indicating critical levels of outbreak [Figure 3]. These patterns suggest that factors other than timing alone must be considered in modeling dengue cases.

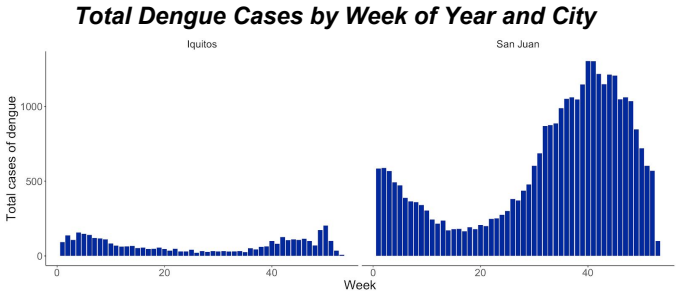
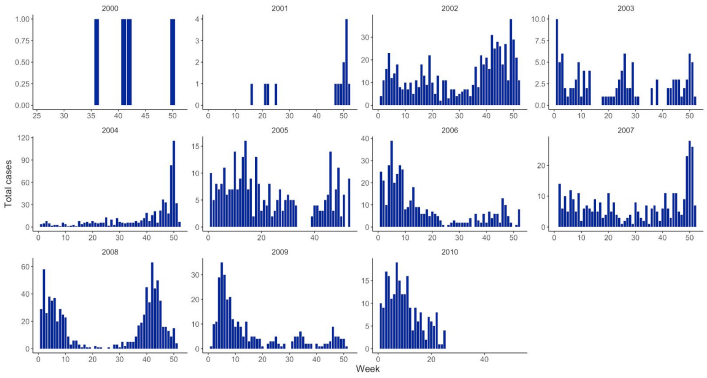


Figure 1: *Total dengue cases by week of year in Iquitos, Peru and San Juan, Puerto Rico.*

Dengue Cases in Iquitos by Year and Week



Dengue Cases in San Juan by Year and Week

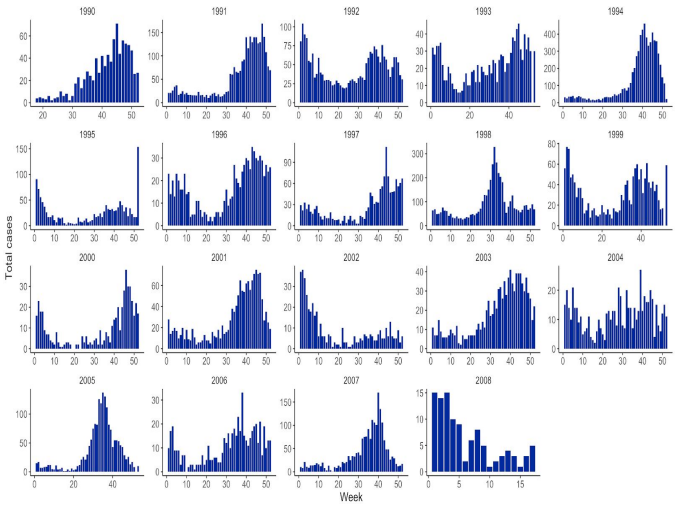
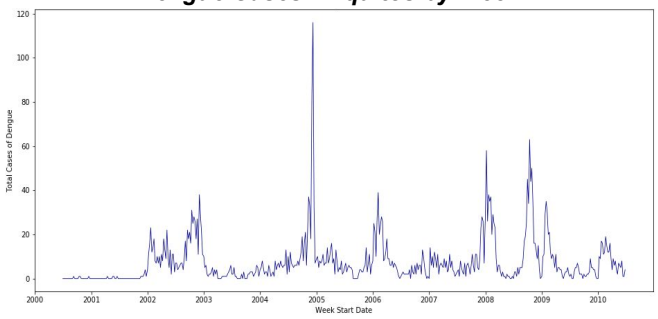


Figure 2: *Dengue cases by year and week in Iquitos, Peru (above) and San Juan, Puerto Rico (below).*

Dengue Cases in Iquitos by Week



Dengue Cases in San Juan by Week

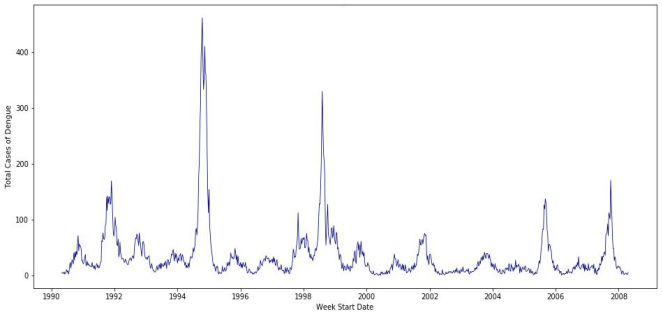


Figure 3: *Weekly dengue cases in Iquitos, Peru and San Juan, Puerto Rico over the entirety of the dataset.*

In exploring related climate factors, we observed that almost 70% of all dengue cases were reported in weeks when the maximum air temperature ranged from 32ºC to 34ºC [Figure 4]. Through further research, we learned that this is because the Aedes genus of mosquito is most active between the temperatures of 30ºC to 35ºC. These mosquitoes cannot feed off blood at temperatures above 36ºC and die at around 40ºC [6]. Since the maximum temperature for both Iquitos and San Juan ranges from 28ºC to 36ºC over the year, the climates of these cities are optimal for mosquito activity.

Dengue Cases by Max Air Temperature (C)

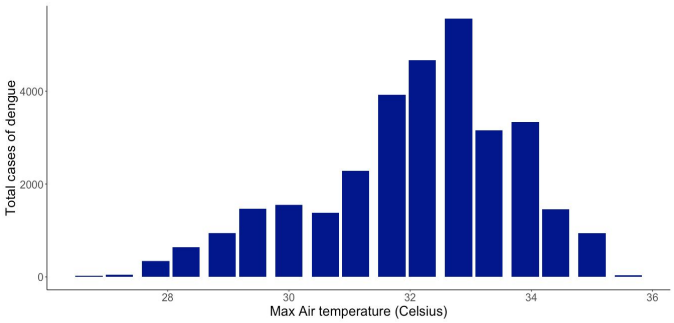


Figure 4: *Total dengue cases by max air temperature.*

A correlation matrix of all variables suggests that some climate variables in the dataset are not independent of one another [Appendix A]. The data was obtained from multiple sources, and hence there are several features reported in duplicate. For instance, the total precipitation is reported once by "Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks (PERSIANN)," and is also reported from NOAA's National Centers for Environmental Prediction (NCEP) Climate Forecast System. The correlation between these two features is of course one, identifying them as redundant. Hence, careful feature selection was necessary prior to modeling the data.

Several preprocessing steps were also required before modeling. Data for the two cities were split so that they could be modeled individually. Redundant variables identified in the correlation matrix, as well as variables with a high number of missing values, were removed through a process of feature selection, and missing values in continuous data were imputed using the mean value of each feature column. These processes reduced the number of climatic variables 16 to 8. For our third model, we also implemented a feature extraction technique, Principal Component Analysis (PCA), on the whole dataset to automate the process of selecting independent features. This is discussed below.

The data was further split into training and testing sets for each city. As this is a time-series problem, the split had to be made such that the models did not use the data from the future to predict the values of the past. Each city's data was therefore split into two continuous parts instead of a random cut. For Iquitos, the data from July 2000 to February 2008 was used as training data, and March 2008 to June 2010 was used for testing. For San Juan, the data from May 1990 to August 2005 was used as training data, and September 2005 to April 2008 were used for testing.

Model Selection

We selected time series models for their ability to handle fluctuations in seasonality, ultimately settling on an Autoregressive Integrated Moving Average (ARIMA) model and Prophet, an open source forecasting model published by Facebook [7]. With only the target variables taken at uniform timesteps as inputs, these univariate methods offered a balance of simplicity and ability to handle the year-over-year and seasonal fluctuations observed in the data.

Given that EDA did reveal evidence of a relationship between climate conditions and Dengue, we also selected a Long Short-term Memory (LSTM) neural network model, which allowed for the integration of both time series and climate features. LSTM is a recurrent neural network that learns from "long term" relationships between factors as well as "short term" trends from new input [8][9]. We tested this multivariate approach on a feature set with and without Principal Component Analysis to see if feature engineering improved our results.

Autoregressive Integrated Moving Average (ARIMA) Model [8]

This univariate approach makes predictions on a target variable using previously recorded (timestamped) measurements of the target variable's values as predictors. ARIMA is a combination of two other time series methodologies: the AR (Auto Regressive) and MA (Moving Average) models. Both AR and MA are based on the assumption that the modeled time series satisfies requirement of stationarity.

In AR, the variable of interest is forecasted using a linear combination of past values of the variable. An AR model of order p is written:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t$$

Where ε_t is white noise [8]. This is similar to a multiple regression with lagged values of yt as predictors.

In MA, the variable of interest is forecasted using past forecast errors in a regression-like model. An MA model of order q is written:

$$t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

where ε_t is white noise. This is a multiple regression where y_t can be thought of as a weighted moving average of the previous several forecast errors.

ARIMA combines AR and MA using differencing to make the time series stationary, where I in ARIMA stands for "integration". The ARIMA model is written:

$$y'_t = c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t$$

where y'_t is the differenced series. The predictors include both lagged values of y_t and lagged errors.

ARIMA requires three parameters:

- p = order of the autoregressive part
- d = degree of first differencing involved
- q = order of the moving average part

Successfully tuning these parameters allows the ARIMA model to operate on conditions of stationarity in applying the linear methods discussed above.

Prophet Forecasting Model [7]

Prophet is a univariate time-series model that takes timestamps and target variables recorded at those timestamps for training. This "black-box" model has three main components: trend, seasonality and holiday effects. These three components are combined to make predictions using the following equation:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t$$

Here, $g(t)$ is the trend function that models non-periodic changes of the time series, $s(t)$ models periodic changes (seasonality), $h(t)$ represents any holiday effects the time series data might contain, and ε_t represents any changes in time series not otherwise accounted for by the model.

This model provides several advantages over other models that we implemented. Unlike the ARIMA model, the time series data does not need to be evenly spaced or cleaned of irregularities; it is robust to missing data and outliers. Also, the model is trained rapidly, allowing for quick updates in future predictions after more data is acquired. It also offers long-term predictions which, although not extremely accurate, provide a general idea of when an outbreak might occur in the future.

Long Short-Term Memory (LSTM) Model [9][10]

LSTM is a type of Recurrent Neural Network model. As a multivariate model, it accounts for the climate data in addition to the weekly number of dengue cases in this particular application. While traditional machine learning models process each observation separately, LSTM "remembers" some information for long periods while integrating new observations. This makes it a good choice for time series modelling.

We trained our LSTM model in a way that if we give it data for seven consecutive weeks, it predicts the number of dengue cases for the subsequent eighth week. However, the LSTM model can be tuned to make predictions further into the future as well.

An advantage of LSTM over ARIMA is that it does not require full retraining as new data is incorporated into the model.

Principal Component Analysis (PCA) with LSTM [11]

PCA is a feature extraction method used to reduce the dimension of the original dataset. Unlike feature selection, in which independent features are chosen manually, PCA creates a new set of independent features that are a linear combination of the originals. The resulting features are ordered in decreasing relevance to the data, with the first feature having the most impact on the dependent variable. One advantage of PCA over feature selection is that it considers the potential relevance of all features, including those that might otherwise be dropped during feature selection. Another benefit of PCA is that the resulting features are independent of each other. However, the dimensionality reduction involved in PCA means that we cannot know which of our original features are the strongest indicators of dengue outbreak per this model, information that may be of interest for ongoing monitoring of climate conditions related to dengue prevention.

Results

Each model reveals both advantages and limitations when applied to our test data sets.

ARIMA

The ARIMA model is highly effective for short-term forecasts with limited input, and predictions become more accurate as new information becomes available. As implemented in this project, ARIMA only predicts one week out and requires "weekly" re-training to be successful, making it less effective for long-term predictions [Figures 5a and 5b].

This could be a useful model in the case of an outbreak, to understand week-over-week needs for medical resources or identify when the number of disease cases has peaked. The test MAE for one-time preditions with ARIMA is 8.14 for Iquitos and 29.69 for San Juan; this error is reduced to 3.80 for Iquitos and 7.19 for San Juan when the model is re-trained weekly.

One-time Test Predictions with ARIMA

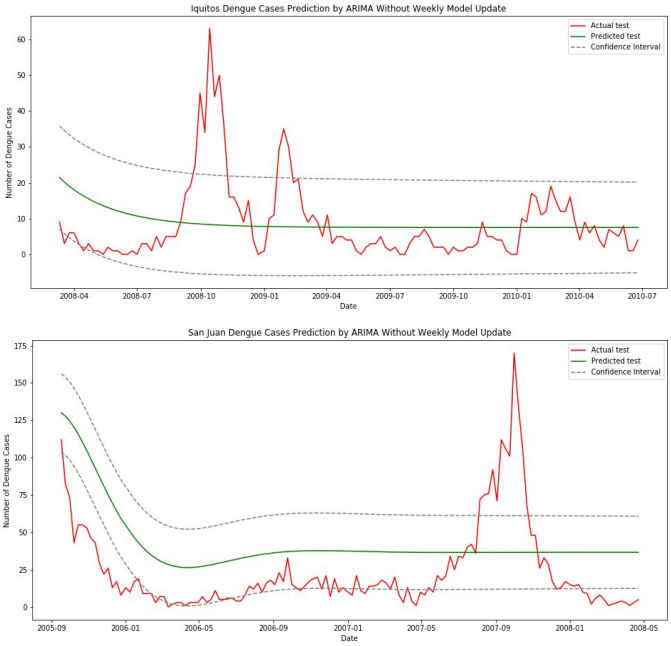


Figure 5a: *ARIMA results with one-time predictions for Iquitos (above) and San Juan (below).*

Test Predictions with ARIMA Re-trained Weekly

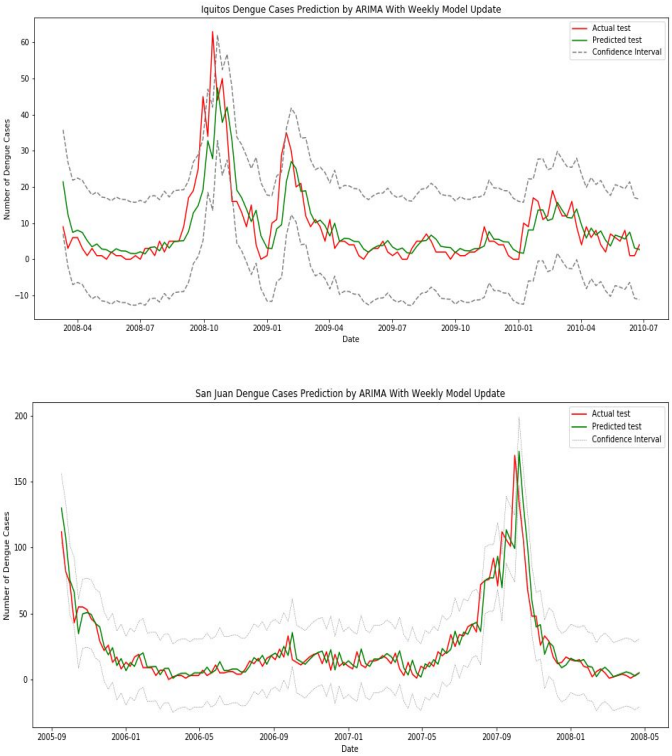


Figure 5b: *ARIMA results with weekly re-training for Iquitos (above) and San Juan (below).*

Prophet

The Prophet model is the least effective of the three, roughly identifying when the number of dengue cases will rise and fall but poorly predicting finer trends in the data or outbreak magnitude [Figure 6]. The Prophet model's MAE is 7.92 for Iquitos and 19.43 for San Juan.

Test Predictions with Prophet

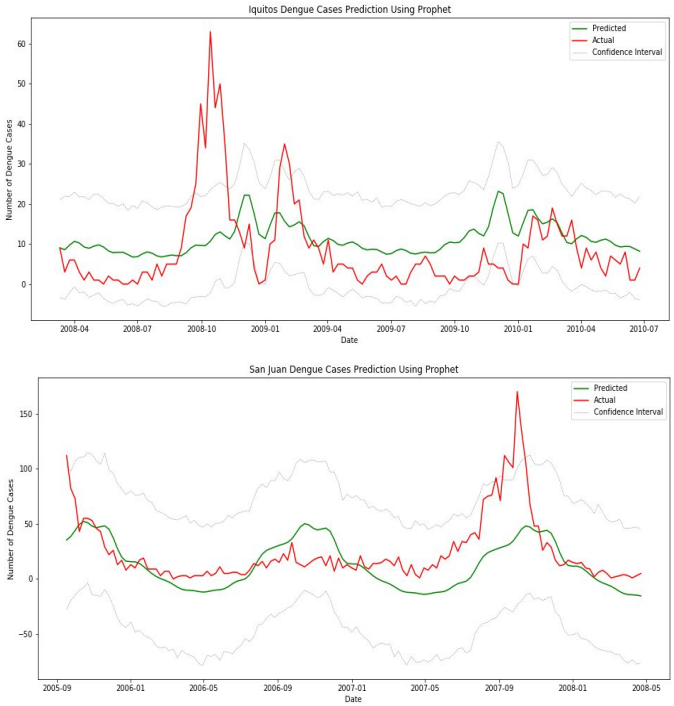


Figure 6: *Prophet results for Iquitos (top image) and San Juan (bottom image).*

Long Short-Term Memory (LSTM)

The LSTM model allows the incorporation of climate factors in addition to the number of dengue cases recorded at each timestep. This model will therefore be able to account for any future changes in climatic conditions [Figure 7]. The MAE of 5.01 for Iquitos is the best result observed for that city. While the MAE of 14.39 for San Juan is not as impressive as the 7.19 observed with ARIMA, LSTM offers a more robust predictive option when considering the influence of climate on dengue.

Principal Component Analysis (PCA) with LSTM

Applying PCA for feature reduction is less accurate in predicting the number of cases when compared to LSTM with manual feature selection. However, LSTM with PCA does accurately predict when an outbreak might occur [Figure 8]. The MAE of LSTM with PCA is 9.08 for Iquitos (using 12 components) and 12.03 for San Juan (using 20 components).

Test Predictions with LSTM

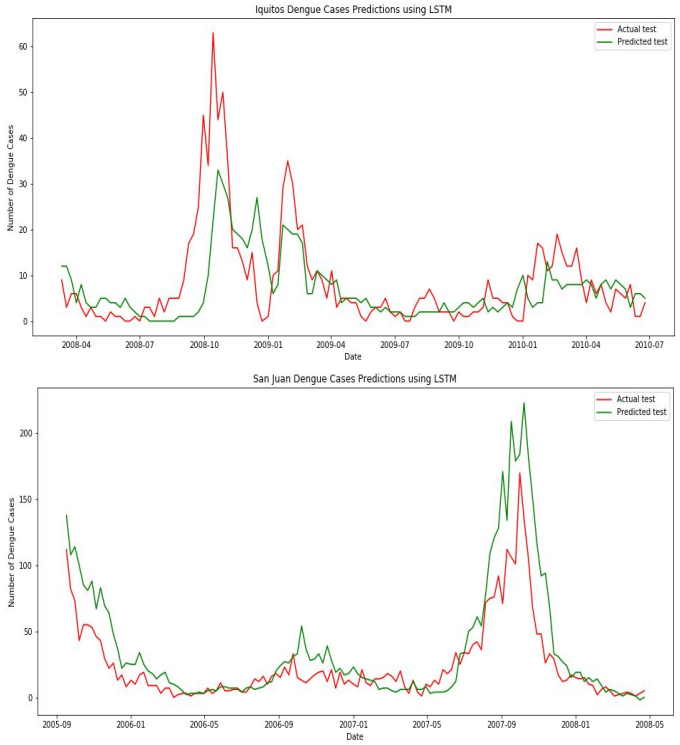


Figure 7: *LSTM results without using PCA for Iquitos (top image) and San Juan (bottom image).*

Test Predictions with LSTM and PCA

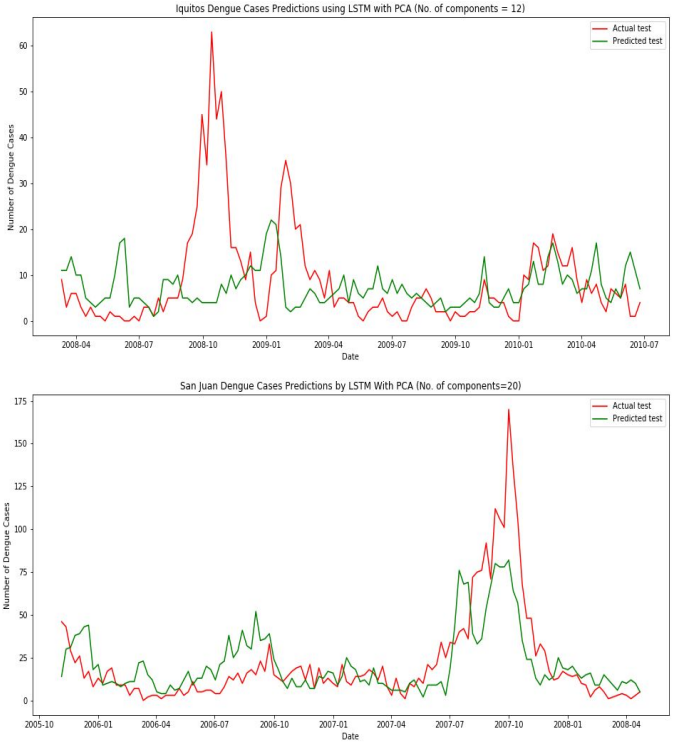


Figure 8: *LSTM results using PCA for Iquitos (top image) and San Juan (bottom image).*

Discussion

Time series model selection involves balancing accuracy with the distance into the future that predictions are desired. These models can be surprisingly accurate with very little input if the interest is in predicting near-term outcomes. The true challenge lies in making accurate long-term predictions.

In the interest of mosquito-borne disease prevention, a long-term prediction identifying *when* an outbreak will likely begin may be sufficient to effect actions (e.g., preventive education, pesticide use) that reduce the scale of the outbreak when it occurs. Accurate short-term models may be useful in predicting scale to meet medical response needs (e.g., supplies and treatment).
Possible next steps include testing these and other time series models for more long-term predictive ability, as well as non-time-series models such as decision trees. Other feature reduction techniques, such as Dynamic Factor Analysis, may perform better than PCA on time series data. These methodologies are worth exploring in future analyses.

Statement of Contributions

Everyone was involved in idea development, EDA, education around time series models, and general discussion of approach. Nitin Kumar Mittal and Smit Kiri implemented the ARIMA, Prophet, and LSTM models. Anuj Anand worked on data pre-processing, implemented PCA, and applied it to the LSTM model. Nicole Reitz and Fenil Shah prepared the presentation and final report, which were finalized with contributions and review from the group.

References

[1] “DengAI: Predicting Disease Spread.” DrivenData, Accessed 22 October 2019, https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/page/80/

[2] “Dengue Fever.” Wikipedia, Accessed 22 October 2019, https://en.wikipedia.org/wiki/Dengue_fever

[3] “Mosquito-borne Disease.” Wikipedia, Accessed 22 October 2019, https://en.wikipedia.org/wiki/Mosquito-borne_disease

[4] Climate change impacts on Dengue virus vectors: https://royalsocietypublishing.org/doi/full/10.1098/rstb.2014.0135

[5] “Dengue Forecasting.” National Oceanic and Atmospheric Administration (NOAA), Accessed 22 October 2019, https://dengueforecasting.noaa.gov/

[6] “Effects of the Environmental Temperature on Aedes aegypti and Aedes albopictus Mosquitoes: A Review”, Accessed 1 Decembeer 2019, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6316560/

[7] ”Prophet.”Facebook Open Source, Accessed 1 December, 2019. https://facebook.github.io/prophet/

[8] Hyndman, R.J., & Athanasopoulos, G. (2018) Forecasting: principles and practice, 2nd edition, OTexts: Melbourne, Australia. Accessed 8 December 2019, OTexts.com/fpp2

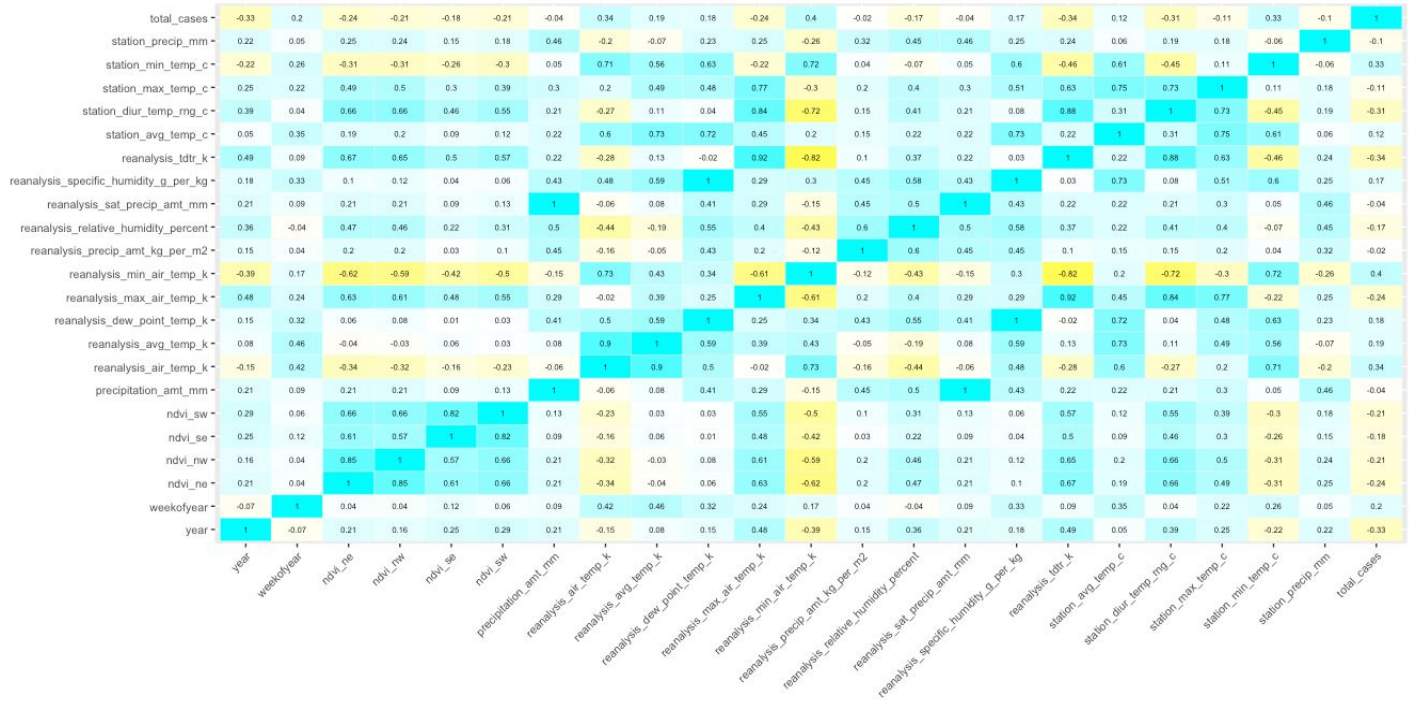
[9] ”An Intro Tutorial for Implementing Long Short-Term Memory Networks (LSTM).” Heartbeat, Accessed 1 December 2019, https://heartbeat.fritz.ai/a-beginners-guide-to-implementing-long-short-term-memory-networks-lstm-eb7a2ff09a27

[10] “How to Reshape Input Data for Long Short-Term Memory Networks in Keras.” Machine Learning Mastery, Accessed 1 December 2019, https://machinelearningmastery.com/reshape-input-data-long-short-term-memory-networks-keras/

[11] “A One-Stop Shop for Principal Component Analysis ” Towards Data Science, Accessed 25 November 2019, https://towardsdatascience.com/a-one-stop-shop-for-principal-component-analysis-5582fb7e0a9c

Appendix

Correlation Matrix



Appendix A: *Correlation matrix showing relationships between variables in the dataset.*