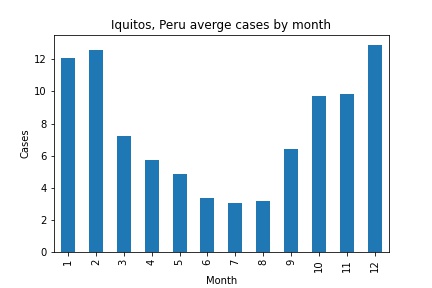
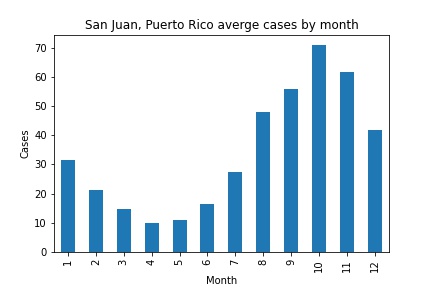
# Dengue Predictions Using Climate Data:

## A Study of Different Modeling Approaches on the Cities of Iquitos, Peru and San Juan, Puerto Rico

Dengue is a mosquito-borne viral infection that can cause severe flu-like symptoms and can develop into a potentially lethal form known as severe dengue. Dengue viruses are principally spread by the female Aedes aegypti mosquito but are also known to be carried by other members of the Aedes genus. “Dengue is common in more than 100 countries worldwide. Forty percent of the world’s population, around 3 billion people, live in areas with a risk of dengue. … Each year, up to 400 million people get infected with dengue. Approximately 100 million people get sick from infection, and 22,000 die from severe dengue.“[[1]](#footnote-0) Many ecological processes are known to play an effect on the dynamics of mosquito populations and the epidemiological impact they have. Many mosquito-borne diseases are known to exhibit seasonal transmission patterns, linked with local environmental variables, like climate variables. Every few years, dengue patterns are punctuated by larger magnitude epidemics, or outbreaks, which are currently unpredictable and are of major concern to populations affected. Other factors are linked with the epidemiological patterns of dengue such as vector population, target population, and a population density which are not considered in this project. A predictive model capable of estimating the magnitude of dengue spread would prove invaluable to the populations afflicted and could afford the opportunity to develop a proactive strategy rather than a reactive one for these populations. This project is a study of different modeling approaches modeling the effect climate data has on the magnitude of dengue cases developed off of the data provided by several US departments.

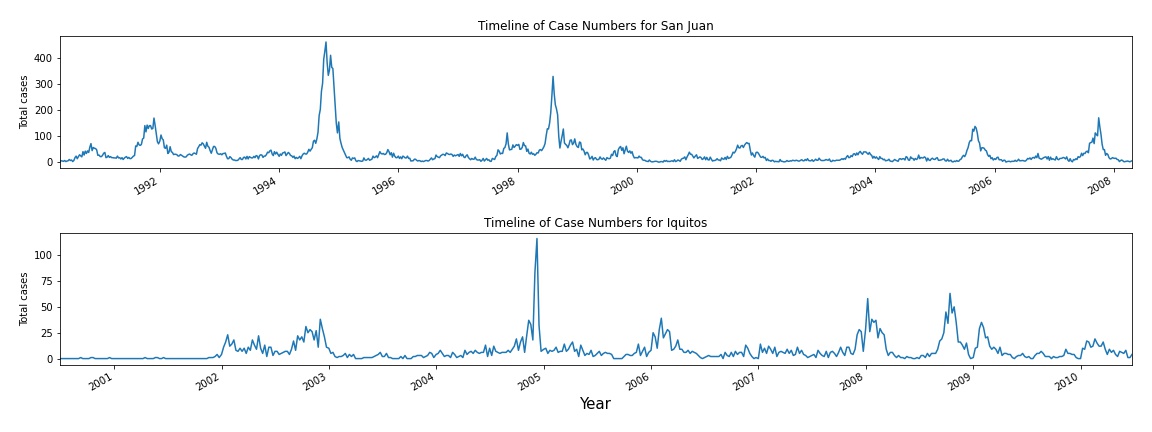
The data provided consists of a compilation of different survey systems recording different features of the climate and geographic surveys for the cities of San Juan, Puerto Rico, and Iquitos, Peru, both tropical geographies. The different US departments (Dept. of Health and Human Services, Dept. of Defense, Dept. of Commerce, and the Dept. of Homeland Security, with support from the Pandemic Prediction and Forecasting Science and Technology Interagency Working Group) initially held a competition open to the public for submissions of a predictive system able to forecast the cases numbers of each city given the climate features included in the surveyed data obtained by the council consisting of these departments. The original competition was held in 2015 and has since resurfaced on DrivenData.org, a hub for data science competitions that have some form of social impact, in 2019. The original competition hosted by the National Science and Technology Council (NSTC) provided the feature data in both CSV and Excel file formats, separate for each city and surveying body. The DrivenData.org competition has taken the step of combining this data into 3 different CSVs, which include one for the feature data that is to be used for training, with an accompanying target data CSV. The third file provided is a feature dataset sampled after the training files that would be used for scoring the performance of the submitted models’ predictions against a withheld testing target dataset not made public.

While DrivenData.org was kind enough to take the step of combining the different data sources, the data within is still not ready for modeling. The features dataset includes several climate features including temperature, humidity, dew point, vegetation index, and precipitation variables from the NSTC with appropriate timestamps associated with when the samples were obtained. Some data sources include a reanalysis of the same data variables provided by different sources though in different units of measure, and there are weeks with missing values for some of these variables, to a varying degree. The month for which each sample was obtained is extracted, and afterward, dummy variables are created to obtain binary features rather than having one categorical feature for model compliance. The units of measure were aligned for the temperature features, which is not a necessary step but it does align the units of similar features for easier interpretation. This is done before dealing with the missing values to help assess imputation performance. A more in-depth imputer was necessary to provide a more realistic pattern in the climate features rather than simple approaches such as filling with the mean of the feature, and an iterative imputation method using a BayesianRidge estimator was utilized. The timestamps of the samples also needed to be reorganized, with many years of samples including a 53rd week of the year as the week the sample was taken. The week labeled as the 53rd week of the year was the first week of the year and every week previous in that year was shifted by one. With the timestamps reordered to reflect the proper week of the year the samples were taken, dummy variables of the monthly values are extracted. The seasonal dynamic of the climate features has some influence on the level of dengue spread, and so an underlying pattern of monthly values could be observed. Some additional preprocessing would be performed for each family of models, as each model has assumptions of the inputted data to properly learn and make predictions.



There is more data made available for San Juan, ranging from 1990 to 2008, though the first and last years are incomplete sample sets. Iquitos has a smaller range of years for which data is available, ranging from 2000 to 2010. Again, only partial data is available for the first and last years. Examining the target variable, there is a stronger underlying seasonal pattern of the weekly reported case numbers for the city of San Juan. Dengue presence seems to worsen during the later parts of the year, with increased numbers in the third and fourth quarters of the year. Iquitos does exhibit some underlying seasonality in the pattern of case numbers reported, though this assumption is weaker and not as prevalent in all years included in the dataset as in the case of San Juan.

The timeline for dengue severity reveals both cities have experienced outbreak periods, and San Juan having higher case values on average than Iquitos. Both cities have periods of several weeks where the case numbers are a significant deviation above the average number of cases. San Juan experienced outbreaks in 1994 to 1995, and in 1998. The worst period of dengue spread in Iquitos occurred in 2005, with smaller outbreak periods in 2008 and 2009 but the order of magnitude of these spikes in case numbers is smaller than those for San Juan.



Upon researching what events may have contributed to these outbreak periods, it was found that in 1998, San Juan was impacted by Hurricane Georges, bringing torrential rainfall in its wake, and the ensuing infrastructural and economic impacts may have had a hand in the increased case numbers. Puerto Rico was dealing with a landfill crisis in 1994 (a problem that persists for many years afterward) where landfills became a public health concern, and garbage runoff impacting potable water sources. No such events were uncovered for Iquitos, but it was found that in 2005, the year corresponding to the largest spike in cases, the Amazon River neighboring Peru experienced one of the worst droughts in recent history. Along with droughts, there are many points in time when the Amazon River has flooded, including in 2008, which aligns with the higher magnitude in case numbers seen in Iquitos. Some of these events are uncontrollable, and all are outside the scope of this project. Instead, this project will focus on the climate data that is provided to stay in line with the more predictable climate features. Of these features, there doesn’t appear to be a strong correlation between any single feature and the case numbers in either city. Interestingly, there seems to be a stronger correlation between the reanalysis temperature features for San Juan than for Iquitos, where these features seem only moderately correlated. Mosquito ecology can be affected by environmental conditions such as the temperature, level of humidity, and amount of precipitation. These two cities seem to have some differences and similarities across their climate features but have a very different magnitude in case numbers.

We looked at precipitation and temperature data as these climate factors are known to affect mosquito populations. The data shows there is not only a difference in the case patterns but also in the climate patterns between the cities. The average air temperature of San Juan ranges from 24 to 29 degrees Celsius for most of the data range, with 2005 and 2006 being slightly colder than the average lowest temperature by a negligible amount. Though Iquitos has a similar range of air temperatures with some more pronounced outliers, the largest difference between the mean temperature and the coldest sample is around 6 degrees Celsius, which wouldn’t be enough to cause major disruption in the mosquito population. Utilizing some statistical approaches, we can hypothesize the true difference between these two cities' average air temperature. Iquitos is located about 250 miles South of the equator, and San Juan is around 1200 miles North, we can use a bootstrap test to run many trials to test whether the mean average temperature of these two cities is almost the same, as observed in the dataset.

To get an idea of how well this data sample is representative of the longstanding climate patterns of these cities, bootstrap replicates are drawn to simulate random samples from the existing dataset. The idea behind this being that the simulated value would remain about the same as the observed sample value if the observed values hadn’t occurred by chance. Drawing bootstrap replicates and utilizing the Central Limit Theorem validated our observations for the mean average air temperature of the cities. Each process resulted in an observed value that was similar, if not identical, to the expected values that were observed in the whole of the dataset for each city, suggesting that the sample data is representative of the similarity in this value on a larger scale for these two cities. The similarity of the mean for the average air temperature readings for each city suggests that this feature is not responsible for the difference in magnitude of case numbers displayed between the cities.

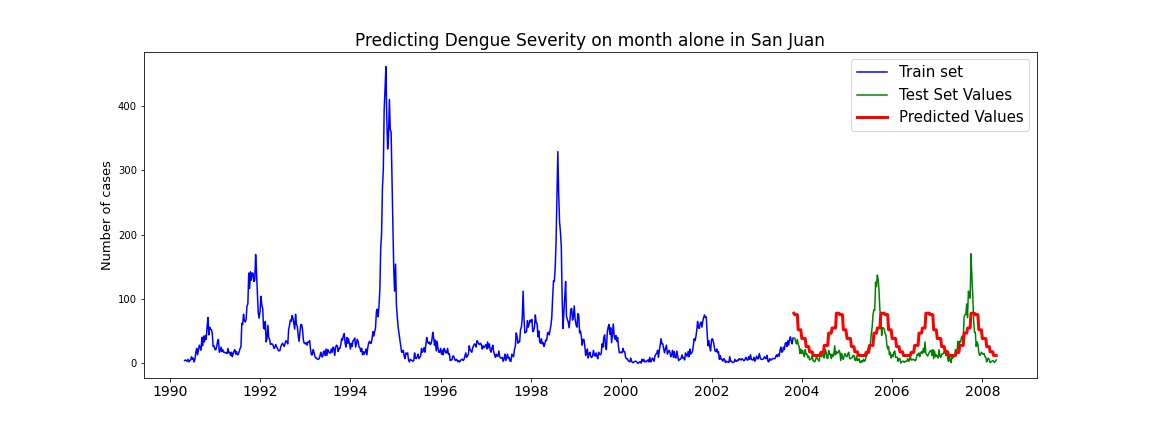
Also using this approach, I looked at the precipitation values as well as standing water is a hotbed of mosquito activity. The Inter-Tropical Convergence Zone, ITCZ, also known as the tropical rain belt, is the condition where the ocean consistently and intensely heats up and evaporates near the equator. Trade winds interact with these convergence zones to carry and shift the convection cycles and “create a persistent band of showers and storms around the Earth’s [equator].”[[2]](#footnote-1) I wanted to look at how these precipitation measurements were different between the two cities. I suspected that with Iquitos being closer to the equator, the ITCZ would bring about higher quantities of rainfall, or possibly more frequent rainfall than in San Juan. The sample data shows that to be a possible likelihood, with the Iquitos dataset having more weeks of heavier rainfall and San Juan having more weeks of lighter rainfall. San Juan also doesn’t reach the same magnitude of precipitation, in millimeters, having an average precipitation value of 40 millimeters below the average of Iquitos. This difference in the precipitation records of the two cities could prove to be of importance, with more precipitation potentially causing more pools of standing water. It would be a fair assumption that increased precipitation patterns could lead to an increase in the mosquito population of the area, yielding a circumstance that could give rise to a systemic increase in dengue spread probability; though, the case numbers don’t align with that hypothesis well since San Juan has overall more cases but less precipitation than Iquitos. A bootstrap test is performed to simulate getting different sample subsets of the precipitation values to hypothesize the true precipitation levels of the cities, and judge where this sample date range is representative of the historic pattern of precipitation for each city. A permutation test of the mean values is performed to determine whether the difference observed in the datasets is representative of the population data for the precipitation values in these two cities. The results showed that the two mean replicate precipitation values were much closer together than shown in our data, suggesting that both locations had close to equal amounts of precipitation on average than what is observed in the sample.

The competition hosted by both DrivenData.org and the original competition hosted by the NSTC, the key performance indicator, the error score that was used to grade the submissions, was decided to be the Mean Absolute Error (MAE). Using the MAE influenced the models to produce flatter predicted values, rather than values that aligned with the variability in the case numbers week by week. Since the internal key goal of different modeling approaches is to reduce a loss by solving a principle problem the model is designed to solve, the predicted results are more encouraged to learn the pattern of case numbers that correspond to the magnitude of case numbers for the majority of the dataset. The issue with this approach is that the model learns to predict values in the neighborhood of the majority of values, weeks not in outbreak periods. As I won’t be participating in this competition, and plan on only using the dataset provided for personal experimentation, the MAE was determined to not be the best scoring metric to use for tuning hyperparameters of the models and grading the models' performance. Other studies in this field have suggested a more appropriate metric to consider would be the Root Mean Squared Error (RMSE). Both metrics provide the error score in relative units, case numbers. The choice of scoring metric will depend on what the goal of modeling the data is.

This project places an emphasis on the prediction accuracy for the weeks of outbreak periods; the RMSE is a more appropriate score to measure. An adaptation of the RMSE score that penalizes under-predictions more than over-predictions that can be used instead is the Root Mean Squared Log Error (RMSLE). Since the predicted values and observed case values are log-transformed, the RMSLE is more resilient to outliers but some models are incompatible with the RMSLE metric when considering the loss function while making predictions, the RMSE is substituted for these situations instead. Other metrics for the model predictions considered include the symmetric Mean Absolute Percentage Error (sMAPE), and the coefficient of determination, R2; however the models present in the accompanying notebooks are not built to optimize these scores. *More information regarding these different metrics can be found in the Dengue\_model\_results document.*

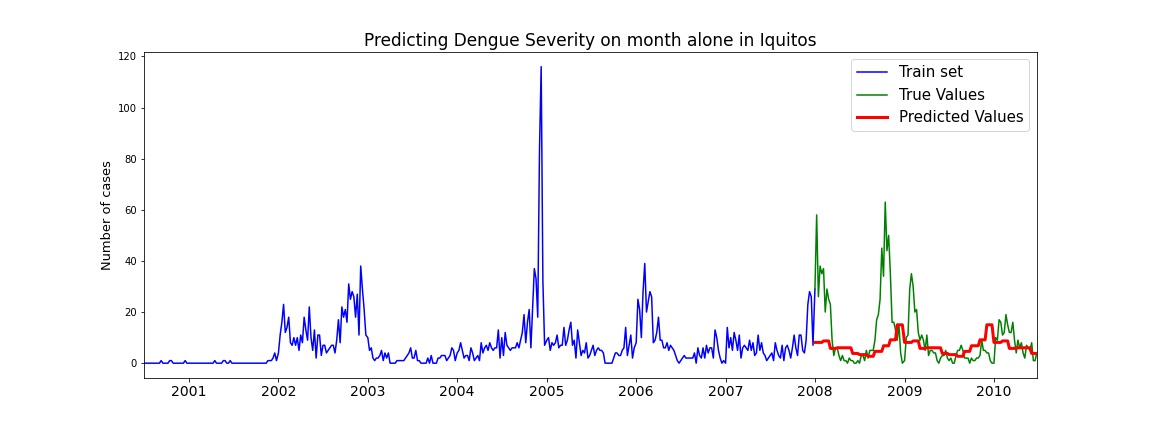
The first approach was a simple ordinary least squares linear regression model. As the case numbers display some form of seasonality, the first model uses the month of sampling to forecast the case numbers. The predictions are cyclical as expected, and the errors for each city contradict each other; the San Juan predictions have a proclivity to over-predict.

**San Juan:**

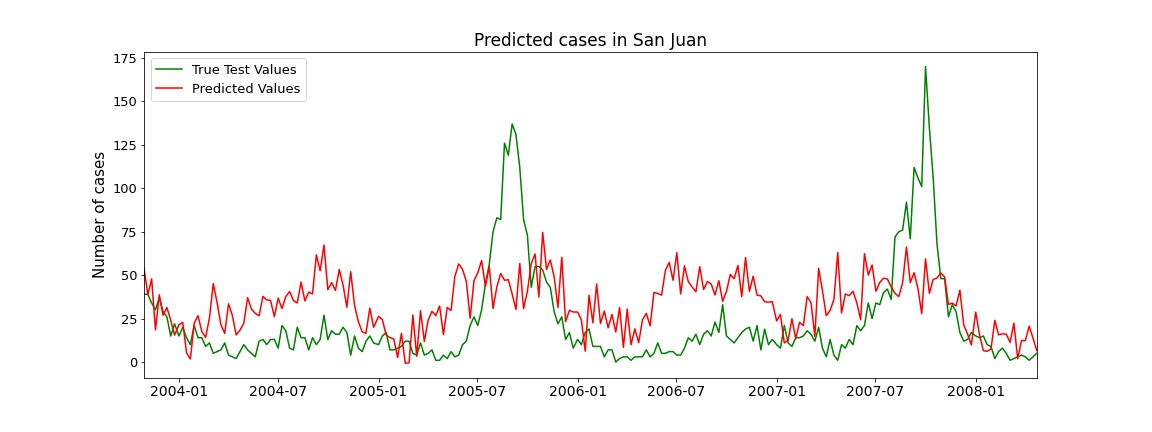


This may be due to the higher case numbers in the training set the model learned on, having a higher magnitude of weekly case numbers and more outbreak periods than what is present in the testing set. Iquitos on the other hand seems to have an almost opposite pattern in the train and test split, having overall smaller weekly case numbers accompanying smaller magnitude outbreak periods in the training set compared to the testing set.

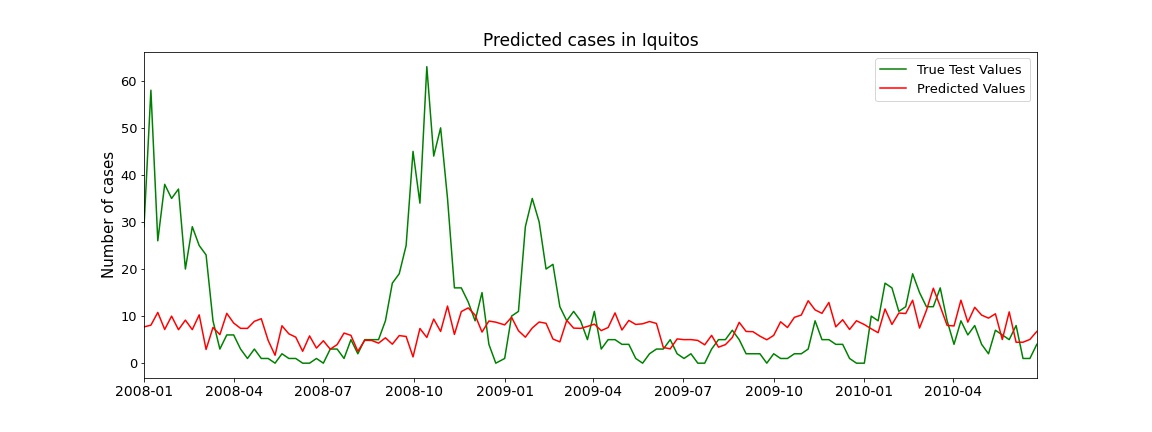
**Iquitos:**



Using the climate features instead, the seasonal pattern of case numbers appears to still be present in San Juan, albeit much noisier and no longer following an exactly cyclical prediction pattern year over year. Though the forecasts still don’t capture the true values of the weekly case numbers in the test set, the MAE of these predictions is smaller than the predictions on months alone, suggesting that climate features do play a role in the dynamic of dengue spread in the city of San Juan. Modeling the climate data for Iquitos sans month variables, the predicted values exhibit less variance, producing a flatter prediction plot. As explained previously, the training set for Iquitos overall consists of smaller case numbers per week for the model to learn from, and the predicted values are on more in line with the non-outbreak week value. The two outbreak periods present in the test set for Iquitos are not captured well by the predictions.

**San Juan:**

**Iquitos:**



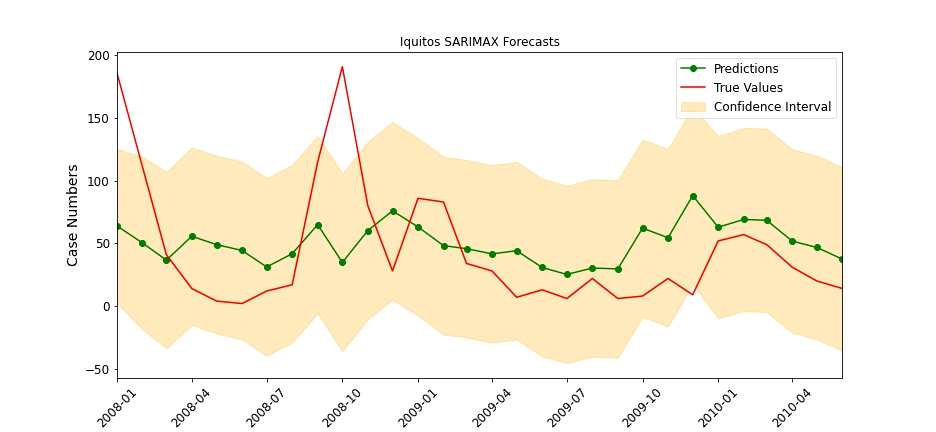
Due to these two periods and the flatter prediction plot, modeling climate data alone produces predicted values and a larger MAE than modeling off of the month variables alone. Simple Linear Regression (including Ridge Regression, Lasso Regression, and ElasticNet models) doesn’t seem to be the correct modeling approach for this time series.

Being a time series, one family of models intended to be used with such a process is the ARIMA (Auto-Regressive Integrated Moving Average) model collection. The competition outline mentions seasonal auto-regressive moving average models as a possible forecasting method to use for making predictions and was a model considered this project. The ARIMA models fit onto the training set of target variables and attempt to define the dynamic of the stochastic process of the dengue spread pattern with the exogenous climate variables in relation to the temporal space in which these samples were recorded. As the processes are interpreted to the time component of the samples, a commonality is required for the frequency across the samples across the sample space. Each year’s samples are recorded on the same day of the month, meaning that the sampling frequency is consistent year after year, but the sampling frequency over all of the dataset is not in a consistent pattern recognizable by the models. For example, each year, the first sample is recorded on the first day of the year and the last sample is recorded on the 24th of December with each sample being taken one week after the previous. This results in samples that are taken in a uniform frequency on a particular day of the week, but are then taken on a different day of the week for the following year depending on which weekday the first of January falls on. Without all of the dataset being sampled on a uniform frequency (same day of the week, exactly one week apart), the models are not able to accommodate this inconsistency.

To suit the requirements of these models, the sum of the number of cases in that month was taken rather than trying to adjust for weekly records. Weekly records would require a lot more processing to not influence the case numbers for the week previous being rolled into the number of cases for the current week; and the symptoms of dengue can take up to two weeks to develop before the afflicted person reports their health status. Tanking the sum of case numbers by month leaves room for error during the first and last weeks of the month but restructuring the samples to be of a common frequency across the entire range of years leaves room for error in every weekly sample, and without more information on how the data was collected and reported, there wouldn’t be a perfect way to split weeks into a uniform frequency.

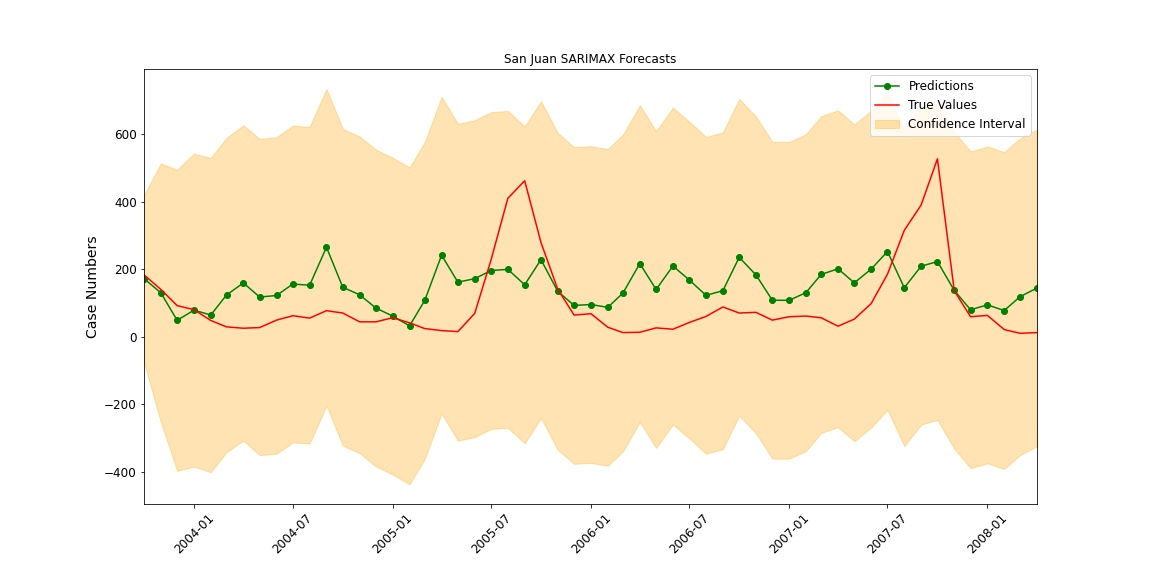
The forecasted values produced by the SARIMAX (Seasonal ARIMA with eXogenous variables) model are the closest to the true case numbers observed for Iquitos of all the models tested. That is to say, the forecasted values capture the variance in the case numbers the best, but that is not to say that these predicted values align with the true case numbers for the Iquitos test set well. In other models tested, the variance in case numbers, being there weren’t many outbreak periods to learn from in the training set, was not captured as well but resulted in better scoring metrics. The metrics reported are not as bad as they appear to be, we need to keep in mind that the RMSE and MAE are error metrics in relative units; the metrics show the error in terms of how the quantity of case numbers the predictions were off by for the summed monthly case numbers. The forecasted values also had a wide confidence band, indicating the confidence interval of the predicted values being within, showing the range the values can fall within and still be considered. Ideally, both the metrics and the confidence band of the predictions are small, leading to a model that is more accurately able to make predictions.

**Iquitos:**

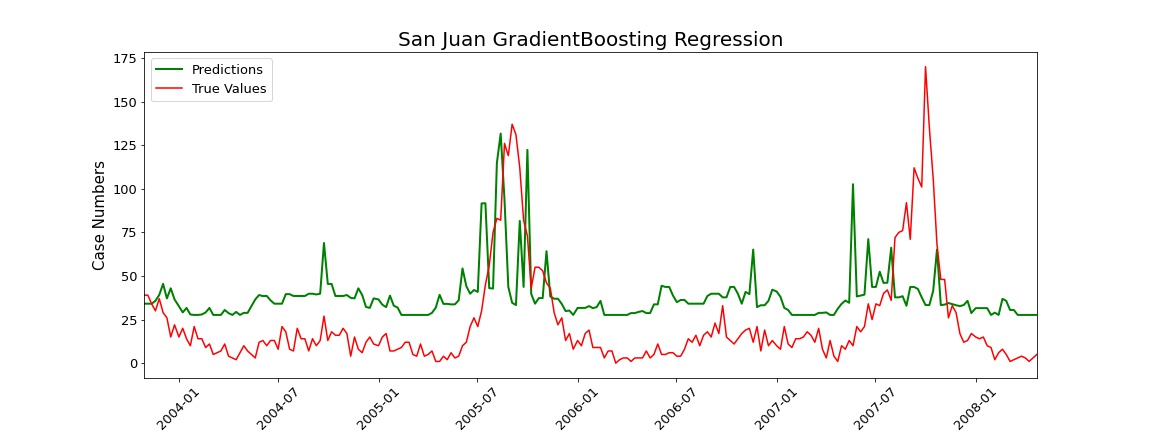


San Juan fairs worse than Iquitos with the SARIMAX model, having much worse metrics as well as predictions that don’t fit the true values well at all. The predictions are overall over-estimating the case numbers for average months and are also under-estimating the outbreak periods case numbers, which again I would consider a consequence of the underlying pattern in the training data. Indeed the predictions don’t seem to account for the variance of the features to meaningful increases in case numbers. Alongside that, the confidence band for San Juan is wider than that of Iquitos; wide enough that predictions well above the peak number of cases in the test set would still be considered possible by the model, and that is reflected in the error scores. I’d like to note that these models were trained on a subset of features manually selected that are understood to have an impact on the mosquito population, and that the auto\_arima functionality provided by the pmdarima library were both cut short because of the limitations of my system, and time complexity required to utilize more features or more complex parameter search by the function. With different feature selections and additional resources, the results of these models may differ. The SARIMAX or other models in the ARIMA family seem more apt at modeling the dengue epidemic in Iquitos than for San Juan.

**San Juan:**

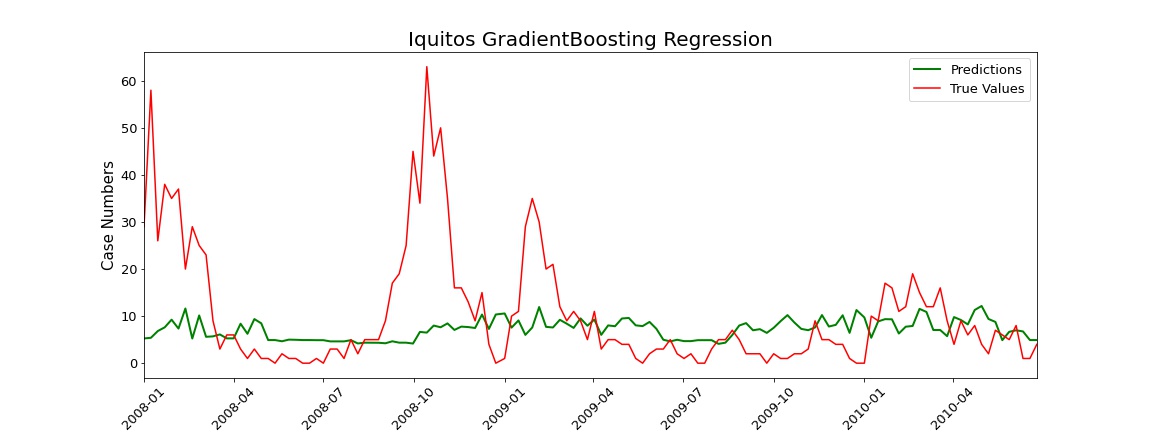


The final family of models incorporates Gradient Boosting. The Gradient Boosting Regressor provided by the sci-kit is able to better capture the variation in the underlying pattern in case numbers for San Juan, while not learning the dengue severity dynamics for the test set weeks for Iquitos. San Juan offers some unique predictions with Gradient Boosting. For the majority of the testing set, the non-outbreak weeks, and offers several spikes during the first outbreak period. These spikes reach the higher case numbers seen in this first outbreak period of the test set, in 2005, but the volatility of these predictions leads to some drops in the predicted case numbers during the worst weeks of dengue cases. The second outbreak period of the test set, from 2007 to 2008, isn’t captured well at all. The predicted values for these weeks fall more in line with the predictions over the non-outbreak periods without reflecting a spike to any extent during this time.

**San Juan:**

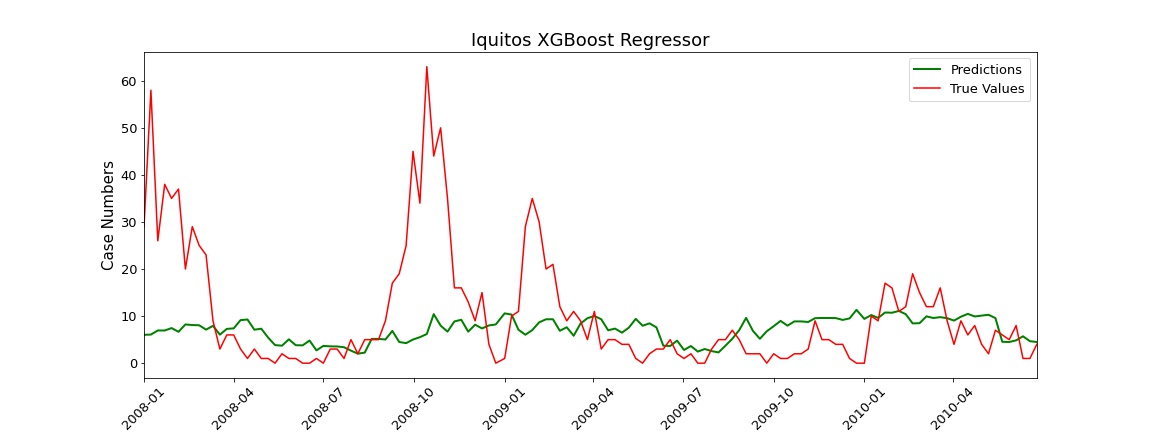
The predictions offered for Iquitos don’t vary at all with the climate dynamics and aren’t representative of the seasonality of the dengue spread. There is evidence the predictions correlate to the pattern in case numbers but there are long spreads of weeks that result in ill-fitting underpreditions or sweeping generalized overpredictions. The predicted values on the Iquitos dataset by the GradientBoosting model are similar to the predictions from the Linear Regression on the non-rolled climate features above.

**Iquitos:**



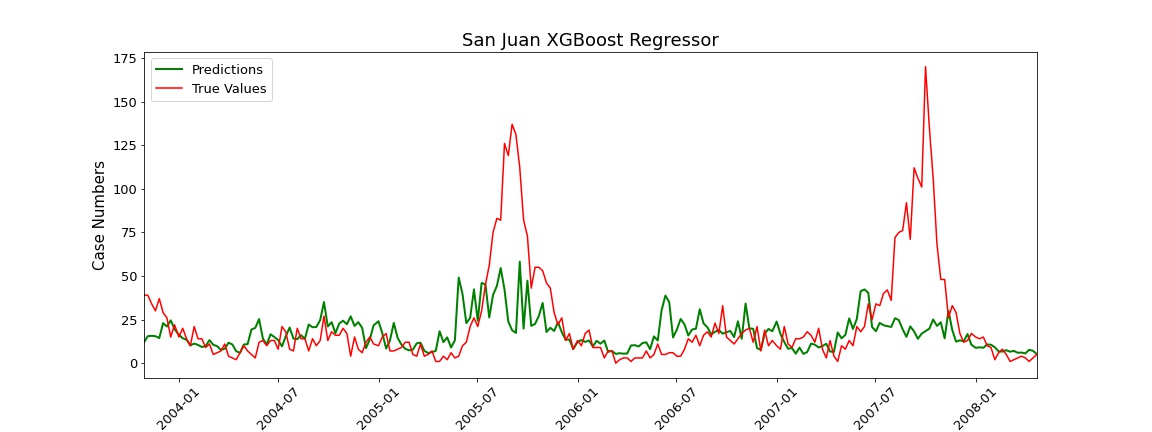
A more extensible approach to the Gradient Boosting algorithm is using the XGBoost library to build parallelized gradient boosting trees. With a larger forest of models learning the data, the resulting predictions provide values with better error metrics, though the prediction accuracy is still left lacking for the outbreak weeks. Iquitos doesn’t benefit in predicting higher case numbers, though there is some boost in performance for predicting the weeks with lower case numbers. The predicted case numbers seem to be less variable than the predictions offered by the base Gradient Boosting Regressor prior, smoothing the predicted trend of dengue severity and biasing towards lower case numbers.

**Iquitos:**



San Juan modeling with XGBoost predicts dramatically different values than those of the GradientBoosting model. The predictions are no longer over-estimating across the whole of the test set and staying more in line with the weekly case numbers of non-outbreak weeks. This does, however, eliminate the strength of the regular GradientBoosting Trees, that is it was able to predict weeks with case numbers in the order of the true first outbreak period; while the MAE and the sMAPE score are improved significantly, the RMSE metrics is unaffected. I’ve found that the XGBoost model improves when a sample weight is specified during the fit to the data, and to the fit of the model, the square root of the case numbers plus an epsilon was provided as the sample weight for XGBoost to consider.

**San Juan:**



A tree-based model seems to perform the best at predicting the cases for the San Juan dataset; GradientBoosting is the modeling approach shown but other tree-based models were also tried and, of the ones tried, seemed to do a decent job at modeling the volatility of the week to week case numbers for San Juan. Other tree-based models tried included the ExtraTrees and RandomForest regression models from the scikit-learn library. The reason Gradient Boosting was chosen in the end was that it offered the best tradeoff of prediction accuracy during the weeks with lower case numbers and error rate during the outbreak periods. XGBoost is an extension of the GradientBoosting model and is influenced by the sample weights used during the fit of the model. Taking the square root of the case numbers with an error term was also a good balance of accuracy overall to error rate during the outbreak periods. Gradient boosting didn’t seem to vary much, but the performance of other models can be augmented by performing some type of feature or component selection.

Many manually selected feature subsets were tried as well as using a cross-validation method for feature selection such as Recursive Feature Elimination, which can be seen in the accompanying notebooks. More approaches can be done to better model this dataset: building an ensemble of models to learn off of the residuals, taking additional exogenous variables into account (social media, population data, hospitalization data for dengue), predicting on a rolling window rather than a weekly basis, and many more. On approach that was tried but not shown in the project includes predicting 4 weeks at a time and refitting the model on the test data afterward iteratively which may prove to be a better approach for different pipelines but didn’t result in significant gains in prediction accuracy for either the outbreak or non-outbreak weeks compared to predicting on the test set en masse.

Whatever approach is used, it seems that external factors such as climate influence the dynamics of the spread of dengue in either city but to be able to understand the interaction of the epidemiology of dengue and the ecosystem for the areas where dengue is endemic, more factors will need to be taken into consideration; population data, vector population, population density, dengue serotypes, economic factors all play a role in the underlying dynamic of dengue spread and severity. As seen in this dataset, some processes are difficult to anticipate that may not lead to a direct impact on the stochastic dengue process but will cause some impact on the populace and can eventually correlate to deviations in the case numbers for dengue.

1. www.cdc.gov/dengue/ [↑](#footnote-ref-0)
2. www.climate.gov/news-features/understanding-climate/annual-migration-tropical-rain-belt [↑](#footnote-ref-1)