University of Victoria

Forecasting Disease Spread

Predicting the spread of Dengue fever

April 16th 2018

Contents

Problem Statement

Our goal in with this project is to predict the spread of Dengue fever in the cities of San Juan, Puerto Rico and Iquitos, Peru. We will be predicting the total_cases for each city, week, year in the test sets. The data ranges from 2008 – 2013 for San Juan, and from 2010 – 2013 for Iquitos. We will be testing and comparing three different models, linear regression, negative binomial regression, and ARIMA, and submitting our results to DrivenData [1]. DrivenData will score our models based on the mean absolute error (MEA). The MEA score is calculating the mean errors across every instance of the test set which is simply the difference of real values compared to predicted values [2]. The lower the score the better the model at predicting the total_cases.

Background

Dengue fever is viral disease spread by the Aedes mosquito that is seen in many tropical countries around the world. There is an estimated 96 million people who suffer from Dengue fever illness each year [3]. There are 4 viruses that cause Dengue fever across the globe, with some countries having all 4. Those that suffer from the Dengue viruses will usually see symptoms such as joint and muscle pain, fever, rash, vomiting, bleeding, and in some cases, even death [3].

Currently there are several vaccines in development with one being recommended by the World Health Organization. However, although the vaccine works in many cases, there are some risks with the vaccine in places where Dengue is not prevalent or the patient has not been infected previously. [4]. Another method of combating Dengue fever is through targeting the Aedes mosquito and attempting to limit breeding habitats [5]. These have not been successful in lowering the numbers of the increasing illnesses of Dengue fever [6].

Although the current methods of combating Dengue fever are not very effective, having an understanding of how and when outbreaks happen is useful information for several countries across the globe. When vaccines are improved or methods for lowering mosquito populations are more successful the data and predictions will be the used extensively. The data may even be used for creating new methods for lowering the number of cases of Dengue worldwide.

Data

The data used in this project was acquired from the DengAI: Predicting Disease Spread competition from the DrivenData website [1]. The basis of this competition and the data given comes from a collective effort of the United States' Department of Health and Human Services, Department of Defense, and Department of Commerce in an effort to predict the next epidemic [7]. This effort started in 2015 and was named the *Dengue Forecasting Project* where the collected data for San Juan, Peurto Rico came in a collaboration of the U.S Centers for Disease Control (DoD) and Prevention and the Peurto Rico

Department of Health while the Iquitos, Peru data came from the DoD in collaboration with Peruvian government, U.S. universities and the Armed Forces Health Surveillance Center [8].

The datasets were presented in two training files: Training Data Labels, and Training Data Features. The "labels" file contained the city, year, weekofyear and total_cases features. The "features" file contained the following features [1]:

- NOAA's Global Historical Climatology Network daily climate data weather station measurements
    - station_max_temp_c – Maximum temperature
    - station_min_temp_c – Minimum temperature
    - station_avg_temp_c – Average temperature
    - station_precip_mm – Total precipitation
    - station_diur_temp_rng_c – Diurnal temperature range
- Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks (PERSIANN) satellite precipitation measurements (0.25x0.25 degree scale)
    - precipitation_amt_mm – Total precipitation
- NOAA's National Centers for Environmental Prediction Climate Forecast System Reanalysis measurements (0.5x0.5 degree scale)
    - reanalysis_sat_precip_amt_mm – Total precipitation
    - reanalysis_dew_point_temp_k – Mean dew point temperature
    - reanalysis_air_temp_k – Mean air temperature
    - reanalysis_relative_humidity_percent – Mean relative humidity
    - reanalysis_specific_humidity_g_per_kg – Mean specific humidity
    - reanalysis_precip_amt_kg_per_m2 – Total precipitation
    - reanalysis_max_air_temp_k – Maximum air temperature
    - reanalysis_min_air_temp_k – Minimum air temperature
    - reanalysis_avg_temp_k – Average air temperature
    - reanalysis_tdtr_k – Diurnal temperature range
- Satellite vegetation - Normalized difference vegetation index (NDVI) - NOAA's CDR Normalized Difference Vegetation Index (0.5x0.5 degree scale) measurements
    - ndvi_se – Pixel southeast of city centroid
    - ndvi_sw – Pixel southwest of city centroid
    - ndvi_ne – Pixel northeast of city centroid
    - ndvi_nw – Pixel northwest of city centroid

The training data ranged from 1990 – 2010 and differs largely between the cities.

Literature Review

We are going to look at several papers on various models that were applied to similar problems as the one we are doing here.

The first paper we looked at is *Time series forecasting using a hybrid ARIMA and neural network model* by G. Peter Zhang [9]. This paper explores the Autoregressive integrated moving average (ARIMA) model as well as artificial neural networks (ANN) for time series forecasting. Because we will be doing time series forecasting for this project we felt this paper could give useful insight into how effective these models are and how they compare to each other and how they can work together.

The ARIMA model is the most widely used time series model due to its statistical properties and well-known methodology [9]. Although the model is flexible in either pure autoregressive or pure moving average, it has its limitations when it comes to being linear. The model is unable to capture non-linear patterns. Due to real-world problems not always being linear in nature, this model may not always be satisfactory. The Box-Jenkins methodology is a common and practical approach to the ARIMA model []. It has three iterative steps, model identification, parameter estimation and diagnostic testing. In the identification step it is necessary for the time series to be stationary. That is, having a mean and autocorrelation constant over time. The parameter estimation step is for parameter estimations to be done with minimal errors. The last diagnostic testing step is to check if the assumption of errors is satisfied.

Along with the ARIMA model, ANN models are emerging and being studied on their effectiveness as a time series model [9]. Compared to the ARIMA model, ANNs are able to perform nonlinear modeling. Most ANNs are done via a single hidden layer feedforward network to achieve approximations of large classes of functions with high accuracy.

When working with data it may not be obvious whether the data is linear or nonlinear, and so a hybrid of the two may create the best results. The paper explores and compares the ARIMA, ANN and hybrid models.

The results show that the hybrid approach did help with reducing forecasted errors [9]. Using a 4x4x1 neural network, the hybrid approach had an improvement in mean squared error over the ARIMA model by 16.13% and the ANN model by 9.89%. The other modifications of the network showed similar improvements using the hybrid model.

Another paper we explore is *Weather as an effective predictor for occurrence of dengue fever in Taiwan* by Wu, Pei-Chih, et al [10]. This paper uses the ARIMA model to predict if weather is an indicator of Dengue fever in Taiwan. This paper's use of an ARIMA model will be comparable to our project and similarities should be seen.

The authors cross-correlated the various variables and cases of dengue fever over the range of -0-5 months of time lag. The data they used in this paper comes from the Taiwan Center for Disease Control and is based around cases in Kaohsuing City from 1988-2003 [10]. Some of the collected data they obtained included the maximum temperature, minimum temperature, humidity and rainfall. They discovered that 1988 and 2002 were the years with the largest scale of Dengue fever outbreaks, with incidences commonly reported after warmer winters and less humid months.

It was found that when cross-correlating the data, weather variables were significantly associated with a rise in dengue fever cases [10]. When applying these variables to the ARIMA model it was found that there was a negative association of dengue fever cases to monthly temperature deviation. There was also a reverse association between total cases and humidity. Both of these findings happened at a 2-month lag period. It will be interesting to see if our findings in San Juan and Iquitos are comparable to these findings and if humidity or temperature will be a significant indicator in dengue fever cases.

A paper by Promprou, S., M. Jaroensutasinee, and K. Jaroensutasinee forecasts dengue fever cases in Thailand using an ARIMA model [11]. In Thailand Dengue fever is considered one of the most important health problems in the country and in its peak year, 1987, it saw cases as high as 325 per 100,000 people. From 2000 – 2004 there were over 100,000 cases and 251 deaths from the illness. This number of cases and death leaves for a desirable way for predicting a potential epidemic.

The data gathered for this study came from across Thailand from 1994 – 2005 and obtained from the Office of Disease Prevention and Control and the Bureau of Epidemiology, Ministry of Public Health,Thailand [11]. The authors used an ARIMA model with the Box-Jenkins approach like the one used in the paper by G. Peter Zhang [11]. A check for stationary was made before the ARIMA model was developed to ensure the model would function properly. They made each observation an expression of a linear function of the previous value of the series (autoregressive parameter) and of a moving average parameter. 144 observations, one for each month of data available, were used in this study. Additional months from January – August 2006 were then used as model evaluation. Based on the data an ARIMA (1,0,1) model was the best fit. They found the model worked well and make predictions for the following year based on this model. The results showed that the regressive forecast curves were comparable to the pattern of actual Dengue fever cases.

The paper *Climatic factors influencing dengue cases in Dhaka city: a model for dengue prediction* by Karim, Md Nazmul, et al. uses linear regression for Dengue prediction in Bangladesh [12]. The collected data from 2000 – 2007 from Bangladesh's Diseases Control room of Directorate General of Health Services and the climatic data from the Meteorological Department at Dhaka. The 2000 – 2007 data was used for the linear regression for the linear regression analysis while they used 2001, 2003, 2005 and 2008 data for retrospective validation of the model. The Average monthly humidity, rainfall, minimum temperature and maximum temperature were the independent variables while the number of

Dengue cases was the dependent variable. Due to there being several independent variables, a multiple linear regression was used.

They found that there were 22,705 cases of dengue fever from 2000 – 2008 with 2002 being the year with the highest total cases at 5851 [12]. 2003 was reported to have the least cases at 450 and with every second year reporting a large increase in cases followed by a year of lower cases. They also found that the most cases happened during the monsoon months. These months had increase in rainfall, humidity and temperature. When plotted it was shown that these months had the highest cases and the post monsoon months showed a steady decline.

They used 3 different models for predicting monthly Dengue cases, one with no time lag, one with 1-month time lag and one with 2-month time lag [12]. Because it can take 7-45 days for a mosquito to go from egg to full adult, a 1 or 2-month lag would likely show the best results for peak mosquito numbers and peak Dengue cases. The results showed that the no lag model was weak at predicting Dengue cases with 8.2% explanatory capability, while the 2-month model showed the highest between predicted and observed cases for their test data.

The last paper we will explore is by Liu, Haibin, et al. and uses Poison regression and negative binomial regression for predicting electric power outages from hurricanes in North and South Carolina [13]. The types of data that were modeled were hurricane-related outages, windspeed, rainfall, land cover type and power system inventories. The purpose of this study was to predict the number of hurricane related outages in an area unit (grid cell or zip code).

The study applied 12 different Poisson and negative binomial regression models applied to grid cell and zip code models and compared the different variables [13]. The results showed that negative binomial regression was seen as more appropriate than the Poisson model. A key difference between the two models could be seen in assigned variance. The assigned variance to the negative binomial model were significantly higher than the Poisson regression model. In particular, the zip code model had the

Poisson model assigned variance at 100 and the negative binomial model at 3,100. However, when the area units had high means then the observed outage counts were more reliable using the Poisson regression model.

Exploratory data analysis

Before we applied models to the datasets, we decided to do a preliminary Exploratory data analysis (EDA) of the data to see if we could find any observations.

The first thing that we did was look at the head of the data. We removed any null values and removed the week_start_date feature as it does not help us at this point. From here we created a variable correlation heat map for each city.

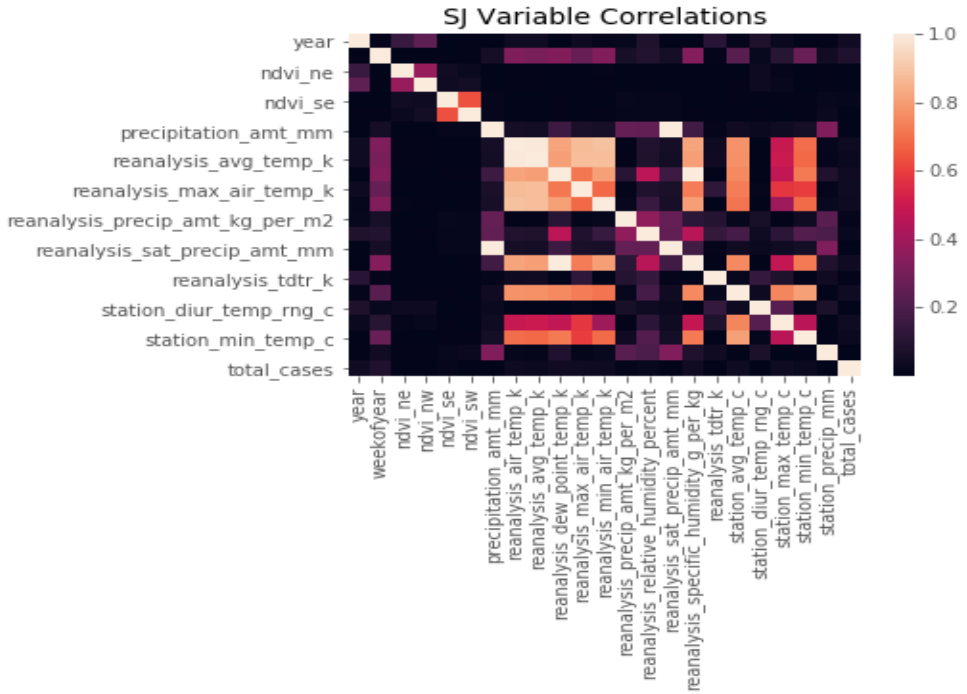


Figure 1. San Juan heatmap

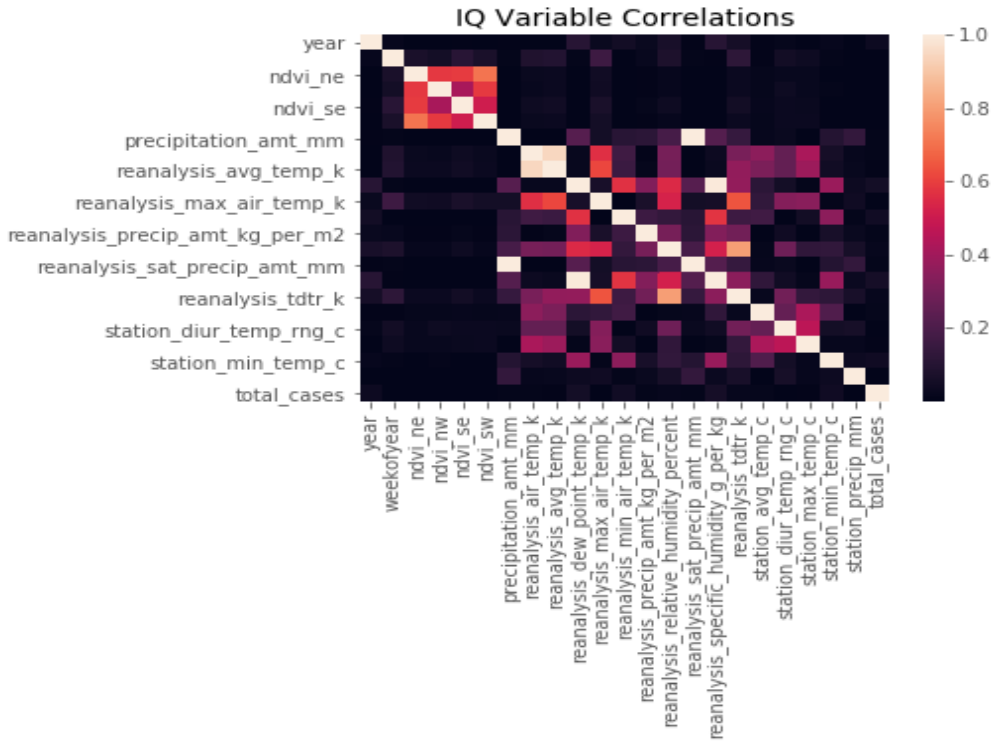


Figure 2. Iquitos heatmap

We can see from the heatmaps that San Juan had strong correlations between temperatures and humidity with weak correlations on everything else. In Iquitos we see that the only relatively strong correlations are between the vegetation indexes, while the rest of the features have little to no correlation with each other.

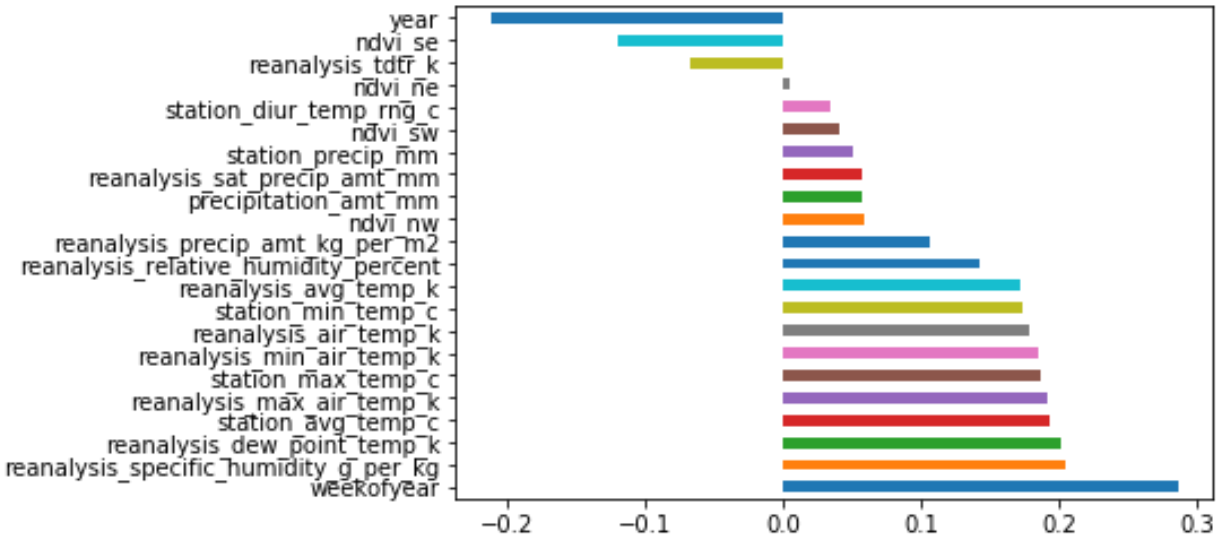


Figure 3. San Juan total_cases correlation

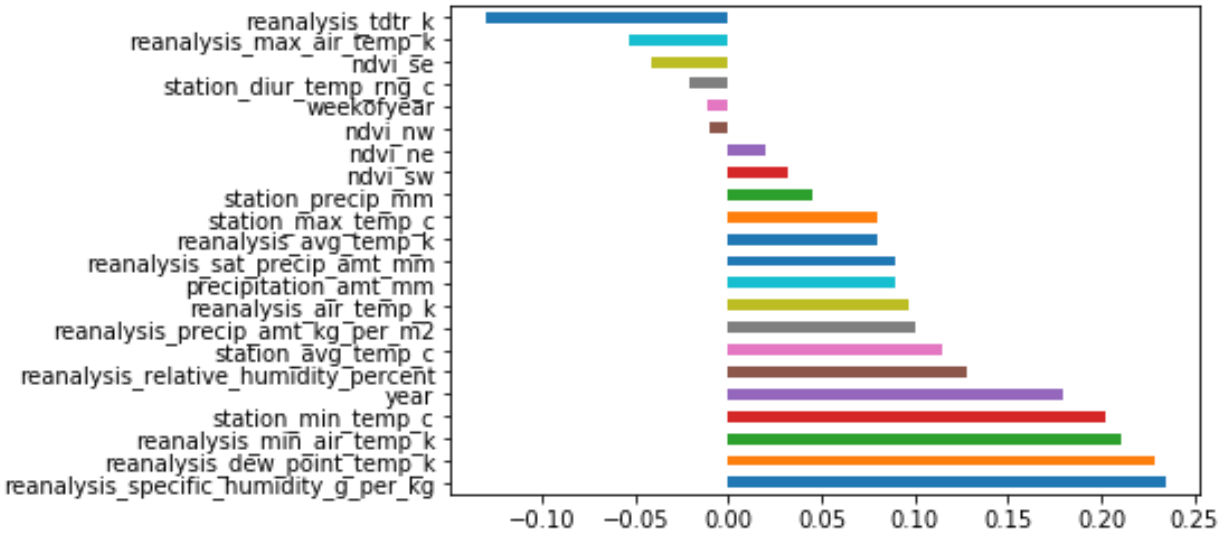


Figure 4. Iquitos total_cases correlation

The bar graphs in Figure 3 and Figure 4 show the correlation between features and Total Cases ordered by negative to positive correlation. Based on the heatmap, it's not surprising to see San Juan have

stronger positive correlations for temperatures with almost every temperature feature having around the same correlation to Total Cases. San Juan also appears to have the strongest correlation be between Total Cases and weekofyear, indicating that the time of year may play a role in how many cases of Dengue fever will be present. However, the year has the strongest negative correlation, which could be due to increased efforts to combat Dengue fever over the years. Other variables factors are interesting, will have to take a look at rolling means correlations when we do Linear Regression.

Iquitos has the strongest correlations of humidity, dew point temperature, and minimum temperatures. In contrast, the maximum temperatures are either weak positive or weak negative correlation to Total Cases. Year is the 5th most correlated feature which might indicate that Dengue fever has been trending upwards over the year.

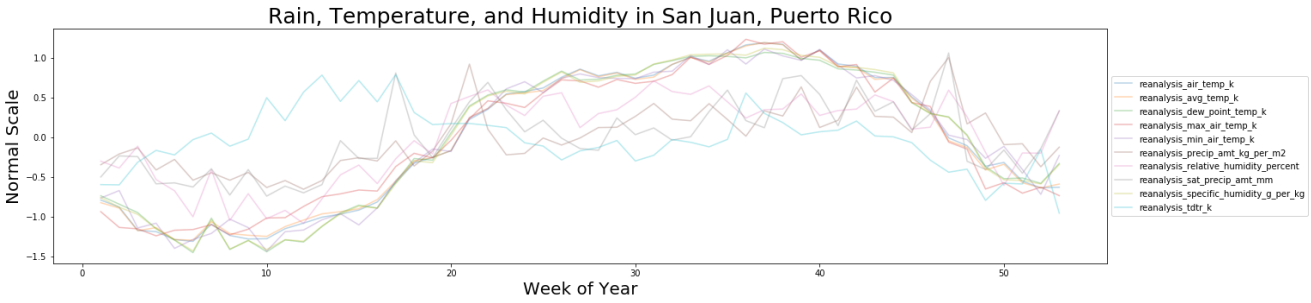


Figure 5. Rain, Temperature and Humidity in San Juan

We can see in Figure 5 there is a very strong trend in SJ from week 10 through to week 37. There is an overall increase in temperature, humidity, and also rain. By week 40 this trend starts to dissipate. It will be interesting to see if total_cases follows a similar trend in those weeks.

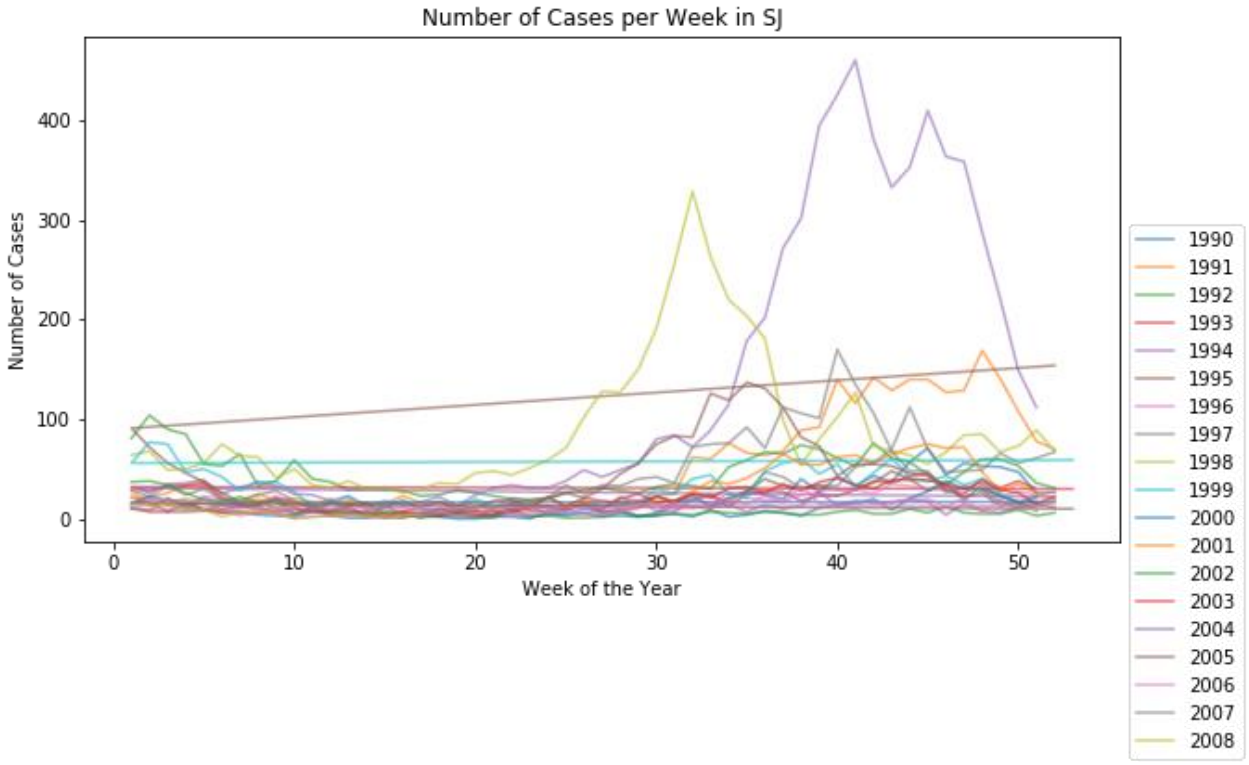


Figure 6. Total cases for San Juan

In San Juan it seems consistent that total_cases starts to rise at around week 25 and starts falling back around week 42. Perhaps the trending perfect conditions we saw before, which start at week 10, take a certain lag in time until they manifest as a reported case in week 28.

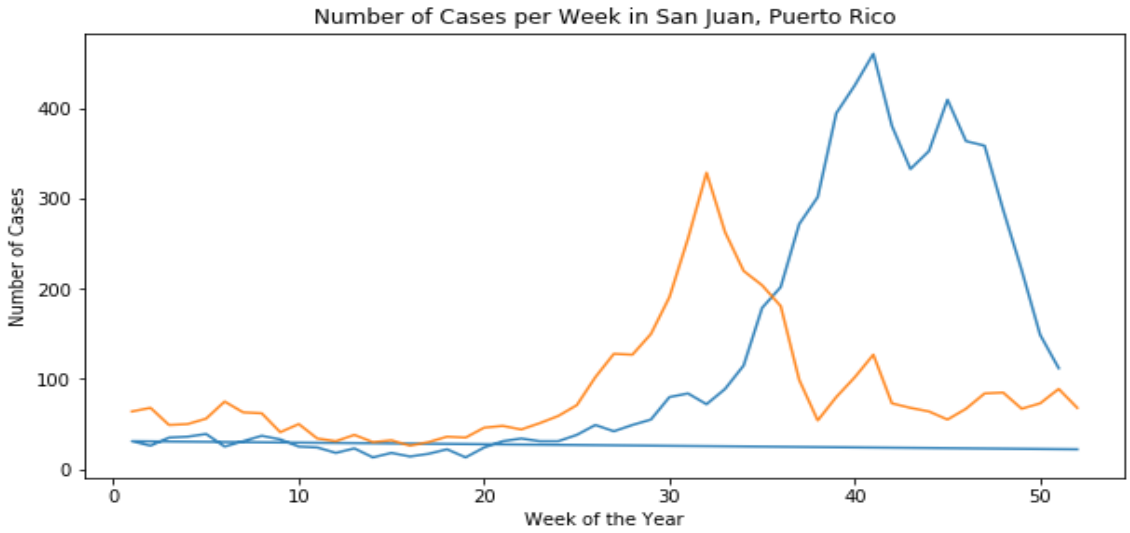


Fig. 7 Peak total cases for San Juan

Here is a closer look at two years with extremely high reported total_cases in San Juan. As we can see in both '94 and '98 the rise starts around week 20. In ‘98 it peaks around week 32 and falls off and in '94 it peeks around week 40.

Differences in San Juan

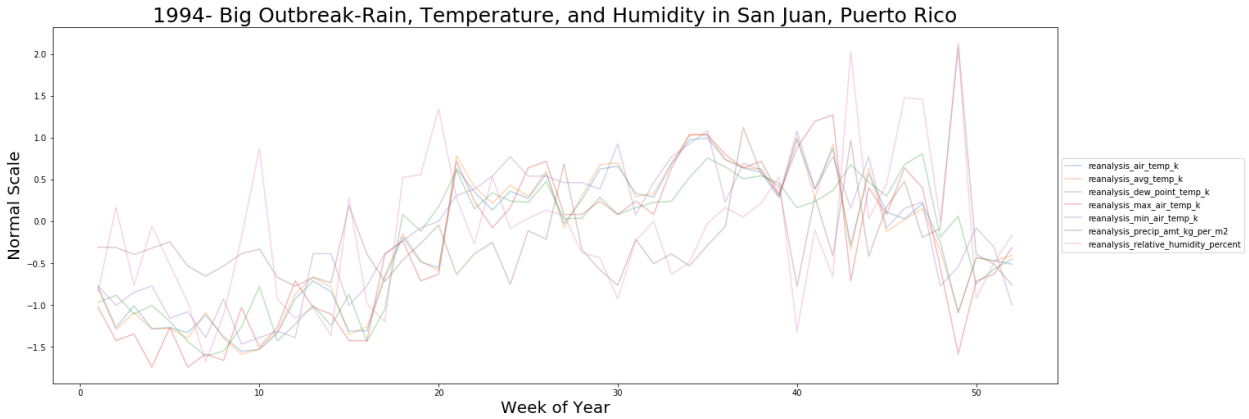


Figure 8. San Juan 1994

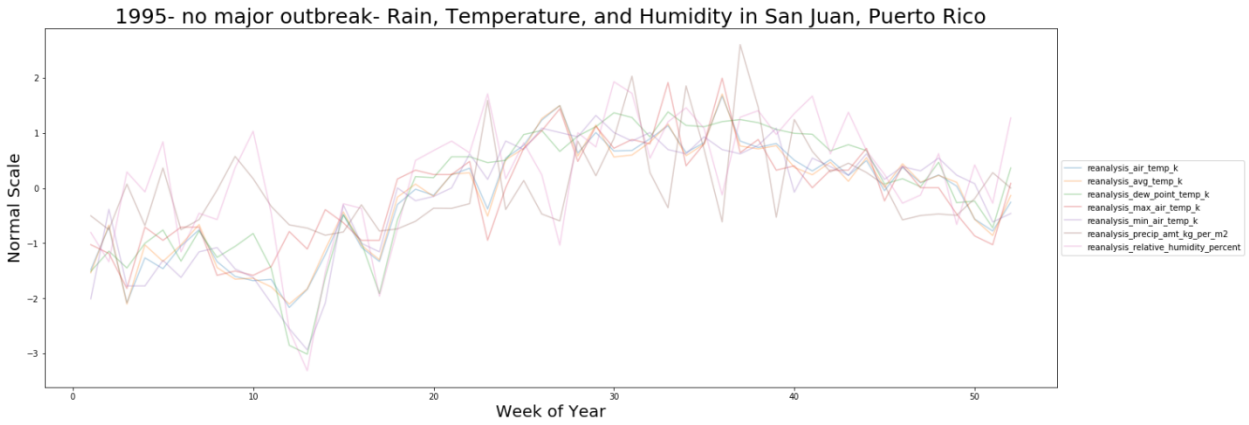


Figure 9. San Juan 1995

We decided to look into San Juan 1994 and see if there were anything in the data that was an indicator for why the total cases spiked in that year. Because of the stronger correlations for temperatures to Total Cases that we found for San Juan, we decided to compare temperatures of 1994 to 1995. The results show very similar patterns between the two years with no notable temperature anomalies in 1994 to indicate this as the only reason for the increase in Dengue fever cases.

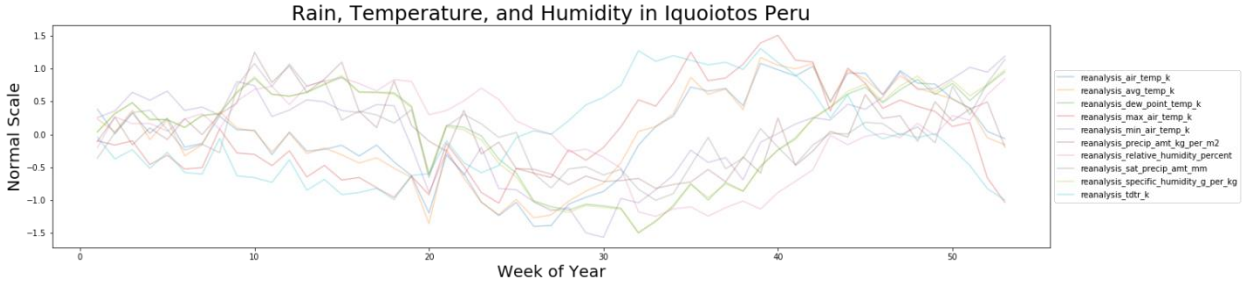


Figure 10. Rain, Temperature and Humidity in Iquitos

In Iquitos the weather factors contributing to the spread of dengue are a bit more convoluted. We can however, see that there is still some sort of trend when temperatures, rain, and humidity all drop between week 10 and 20 then stagnate until week 30, where they all begin to dramatically rise until the end of the year.

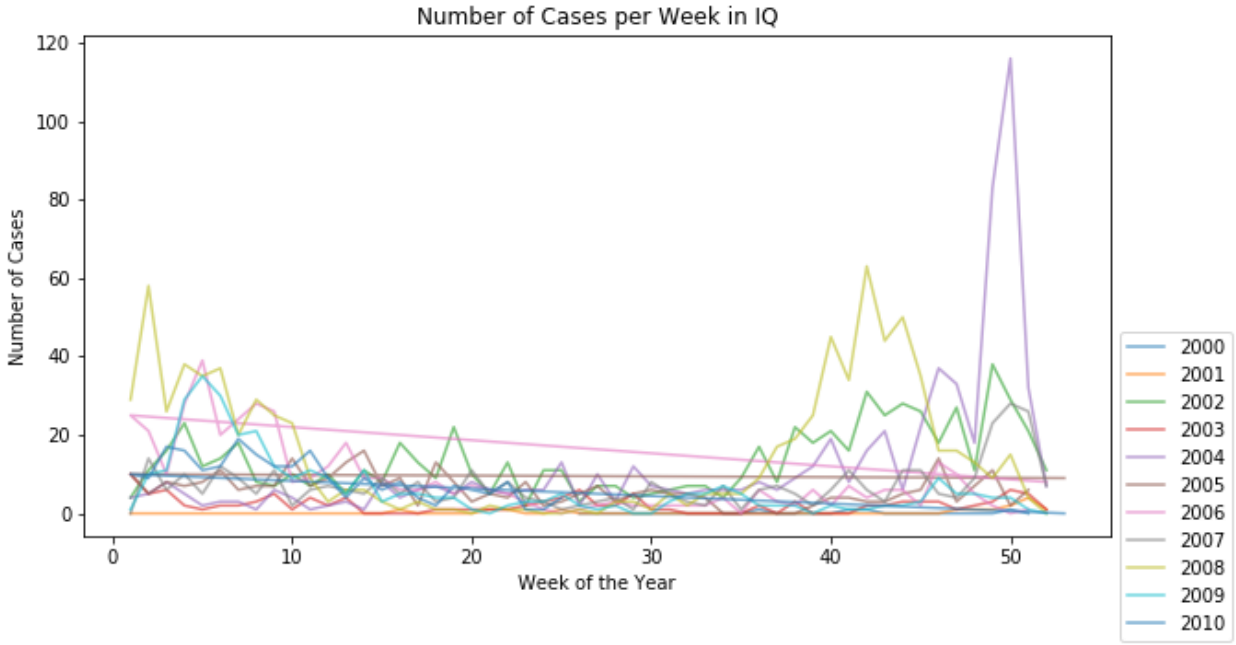


Figure 11. Total cases in Iquitos

The for total_case in Iquitos is a little more erratic than in San Juan. We can see an overall pattern where total_cases start off higher in the beginning of the year and falls off by week 20 where it remains down until it starts growing again around week 38. This trend seems similar to the weather trend we saw earlier.

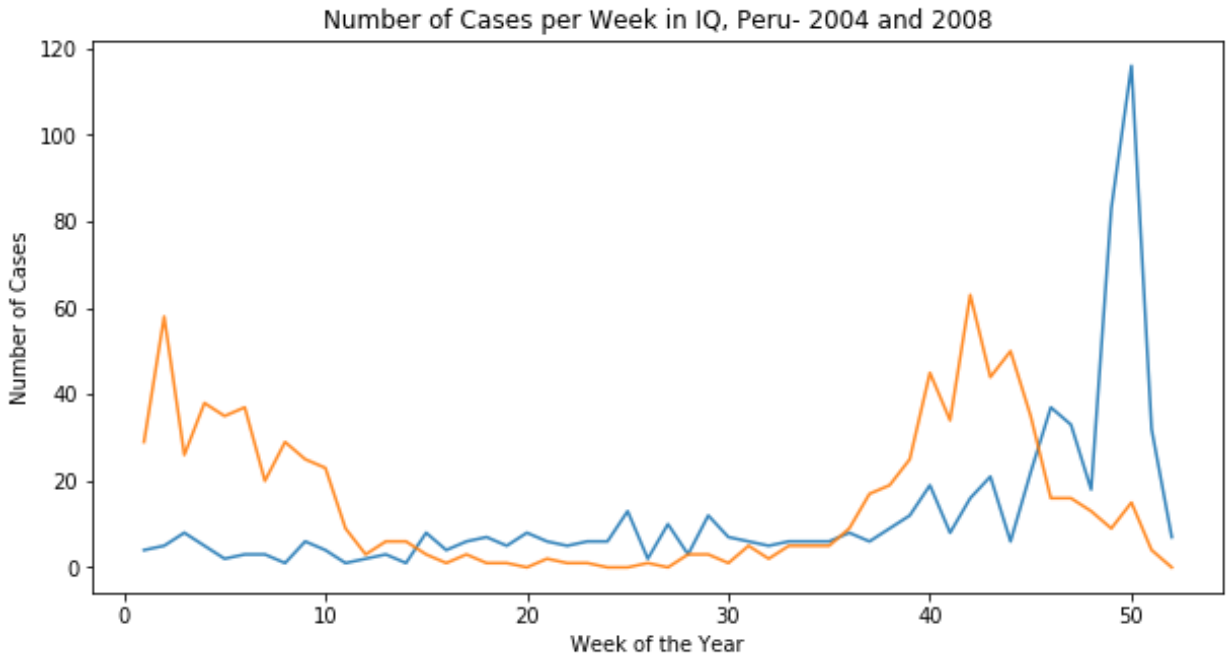


Figure 12. Peak total cases Iquitos

Looking at the two outbreaks in 2004 and 2008 in Iquitos we see rise starts around week 35 in 2008 and around week 50 in 2004 where there is a very sharp peak before quickly subsiding.

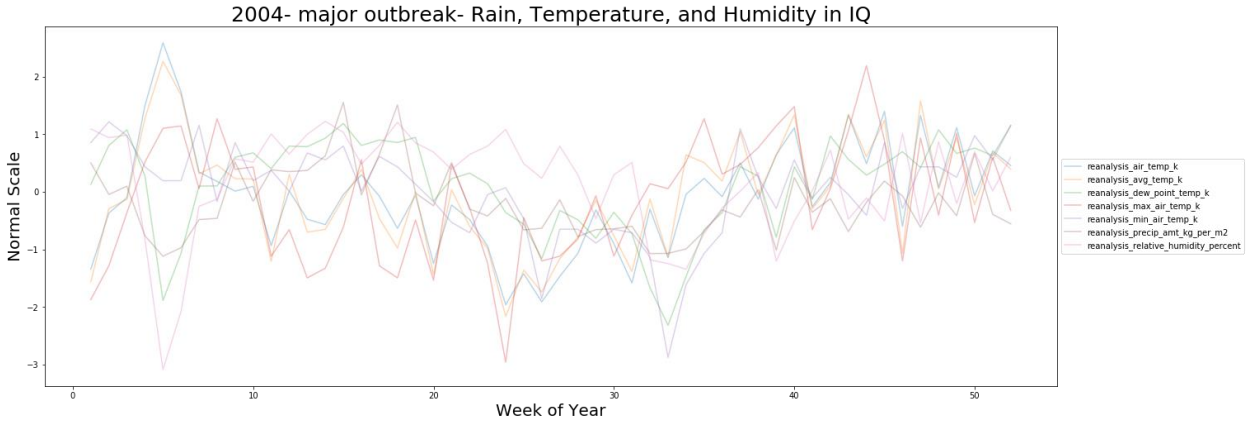


Figure 13. Iquitos 2004

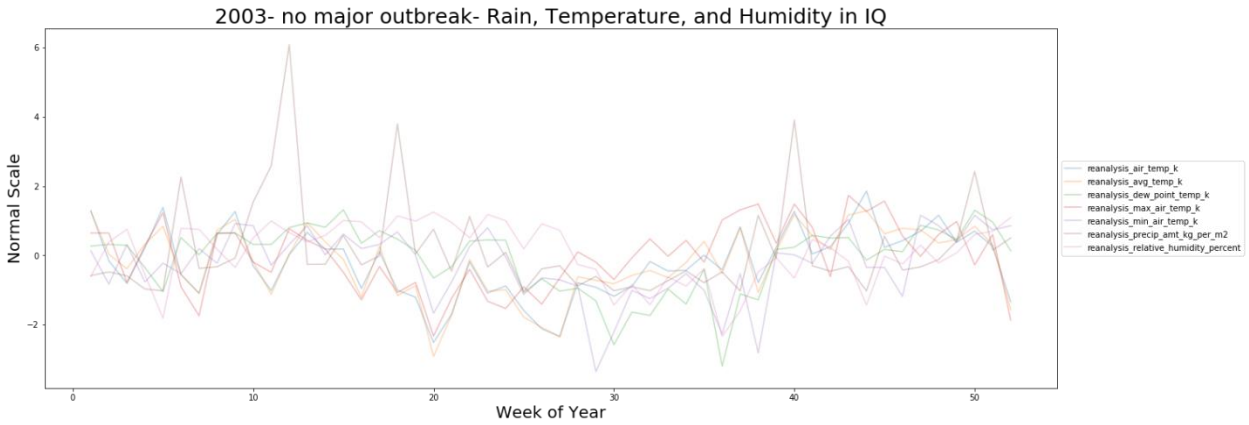


Figure 14. Iquitos 2003

We applied the same tactic for Iquitos and compared the humidity of 2003 to 2004. The graph shows that both years had similar patterns and in areas with large differences. Weather patterns stay mostly consistent the entire year of 2003 when there was no major outbreak. However, they are also relatively all lower than in 2004. Perhaps it is the inconsistence and fluctuations in weeks leading up to week 40 in 2004 that caused such high numbers.

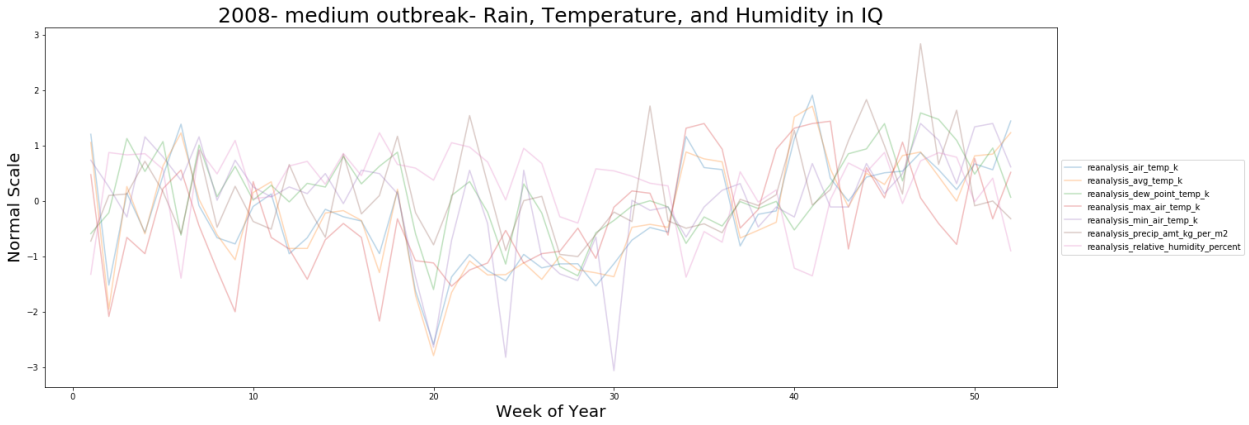


Figure 15. Iquitos 2008

In the two years where there were big outbreaks, 2004 and 2008, we see comparable graphs. There is a lot of fluctuation leading up to week 30 until the variables begin growing at a similar pace and end up moving together. It will be interesting to try modeling the seasonality of dengue spread with San Juan as it seems there is a strong trend to the dips and peaks of the disease no matter if it's an outbreak year or normal year. Iquitos might be harder to model based on this entry analysis. There are less outbreaks and less of a trend for how the disease spreads.

Results

ARIMA Model

The ARIMA model stands for autoregressive integrated moving average [14]. The model uses autocorrelations which are measure how strong values are in specified periods apart are to each other over a lag period [14]. It also uses moving averages which relates what happens in a certain period of times to random errors that occurred previously [14]. Mixing these models together generally leaves for more accurate forecasts. The first model we built was the ARIMA to try and address the time-series and data aspect of the series. We simply wanted to see if, just using the total cases of dengue, we could identify an overall pattern through the years which was strong enough to create a reliable trend prediction and build monthly or seasonal changes on top of that. A quick decomposition of the time series date showed that there is no clear trend to the spread of dengue through the years, but there is seasonality to the disease, meaning that within any given year the monthly course of how total cases are affected is very steady. Furthermore, most dengue cases in outbreak years could be clearly observed in the residuals plot which showed outliers from the over-all trend.

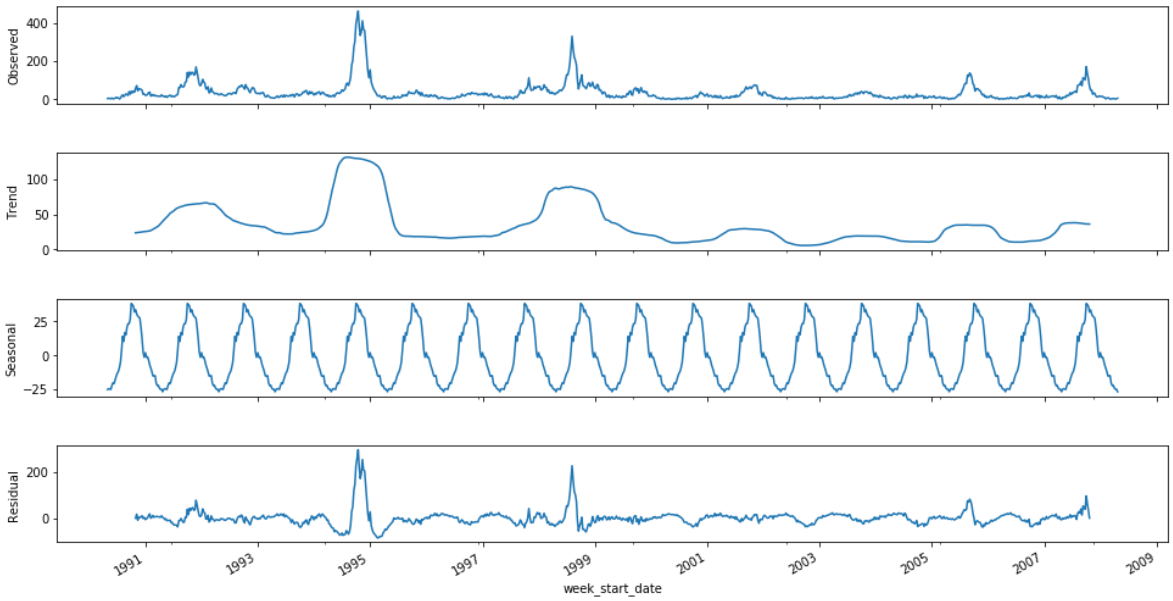


Figure 16.

Decomposition of time series for San Juan, Puerto Rico.

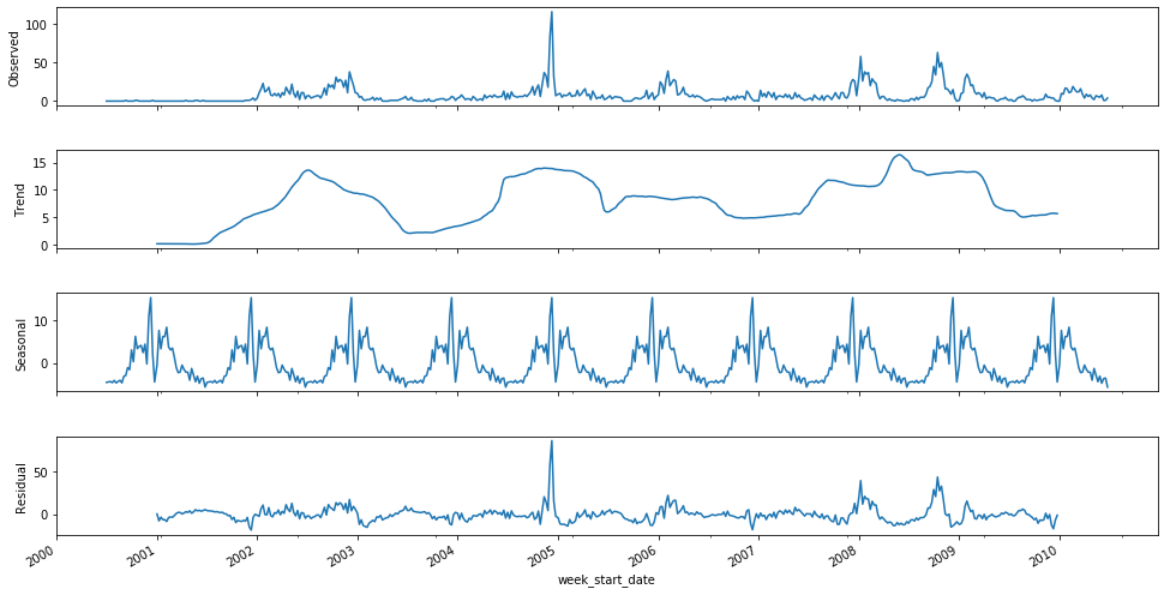


Figure 17. Decomposition of time series for San Juan Puerto Rico

In order to achieve a more stationary time series, that is one with a stable or constant mean and variance, we used differencing. We specifically used seasonal differencing, taking the difference of total cases between a current year and the previous year, and testing stationarity with the difference. This one degree of differencing gave us a test-Statistic lower than the 1% Critical Value, leading us to believe that we were ready to start modeling.

```
Results of Dickey-Fuller Test:
Test Statistic                  -1.046406e+01
p-value                          1.335837e-18
#Lags Used                       1.200000e+01
Number of Observations Used      8.700000e+02
Critical Value (1%)             -3.437889e+00
Critical Value (5%)             -2.864868e+00
Critical Value (10%)            -2.568542e+00
```

Figure 18. Test statistic after one degree of differencing for San Juan

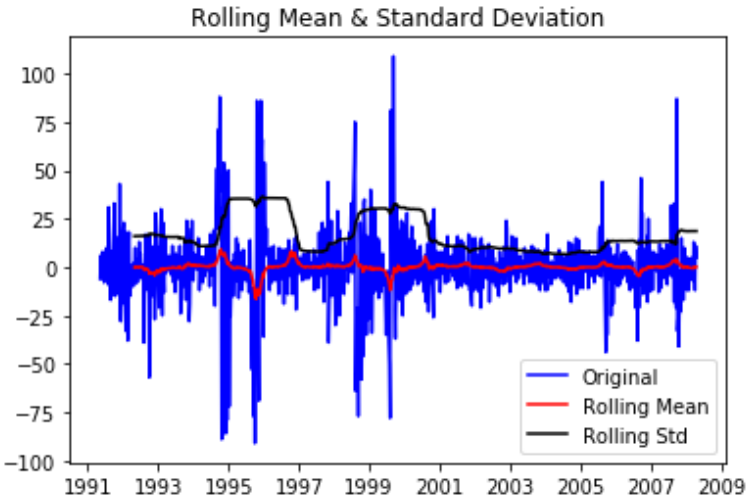
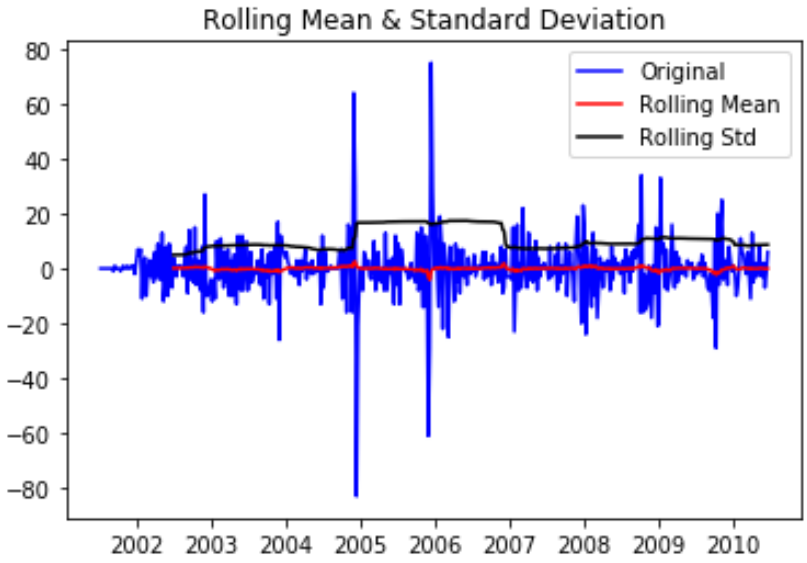


Figure 19. Mean and Variance after first degree differencing for San Juan

```
      Results of Dickey-Fuller Test:
Test Statistic                -9.276254e+00
p-value                        1.286610e-15
#Lags Used                     1.100000e+01
Number of Observations Used    4.550000e+02
Critical Value (1%)           -3.444804e+00
Critical Value (5%)           -2.867913e+00
Critical Value (10%)          -2.570165e+00
```

Figure 20. Test statistic after one degree of differencing for Iquitos

Figure 21. Mean and Variance after first degree differencing for San Juan



The ARIMA model has three parameters : p,d,q. The p value is responsible for determining how many AR windows we implement. The d is the degree of differencing. And q is the number of MA windows we use. To determine the values for and q we need to see the auto correlation and partial auto correlation the total cases exhibit with themselves during different time lags.

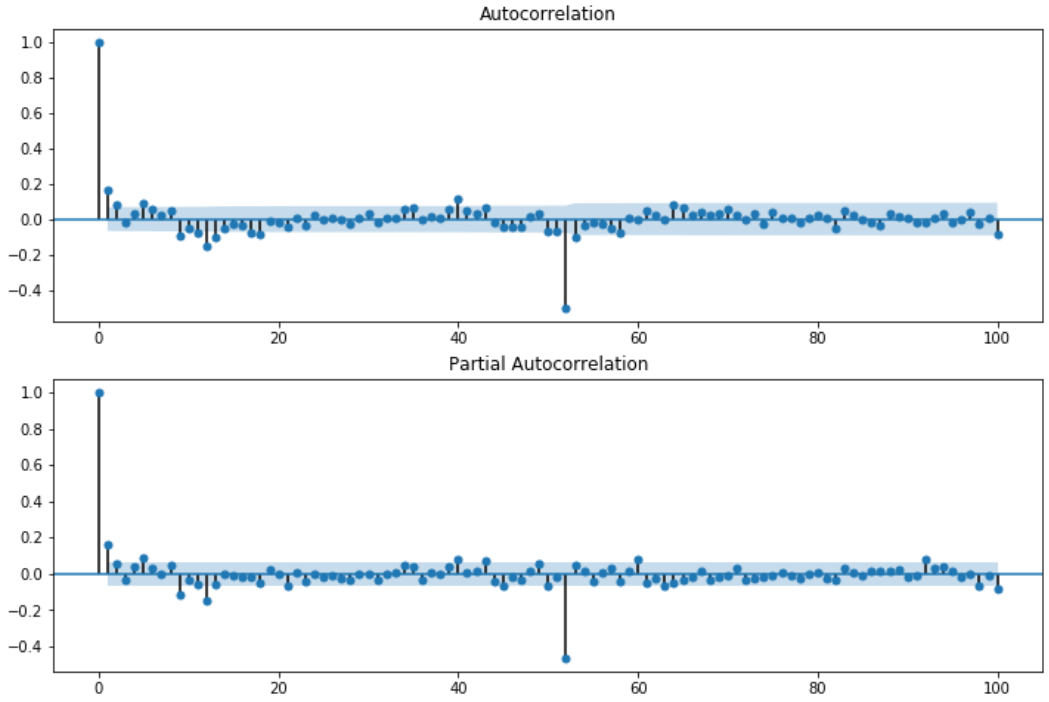


Figure 22. Auto and Partial Auto Correlation for SJ

As we can see from figure 8: The total cases have a significant autocorrelation after two time lags, and a significant partial autocorrelation after one time lag. This leads us to use a p value of 2 and q value of 1 for a ARIMA (2,1,1) model. As we can see from figure 9: The total cases have a significant autocorrelation and partial autocorelation after two time lags.. This leads us to use a p value of 2 and q value of 1 for a ARIMA (2,1,2) model. In both cities, we see a significant negative autocorrelation and partial auto correlation around week 52 and that is because we see a year full lags, mean the same weeks number the following year are related, this could be due to seasonal trend and we will disregard these correlations when building our ARIMA model.

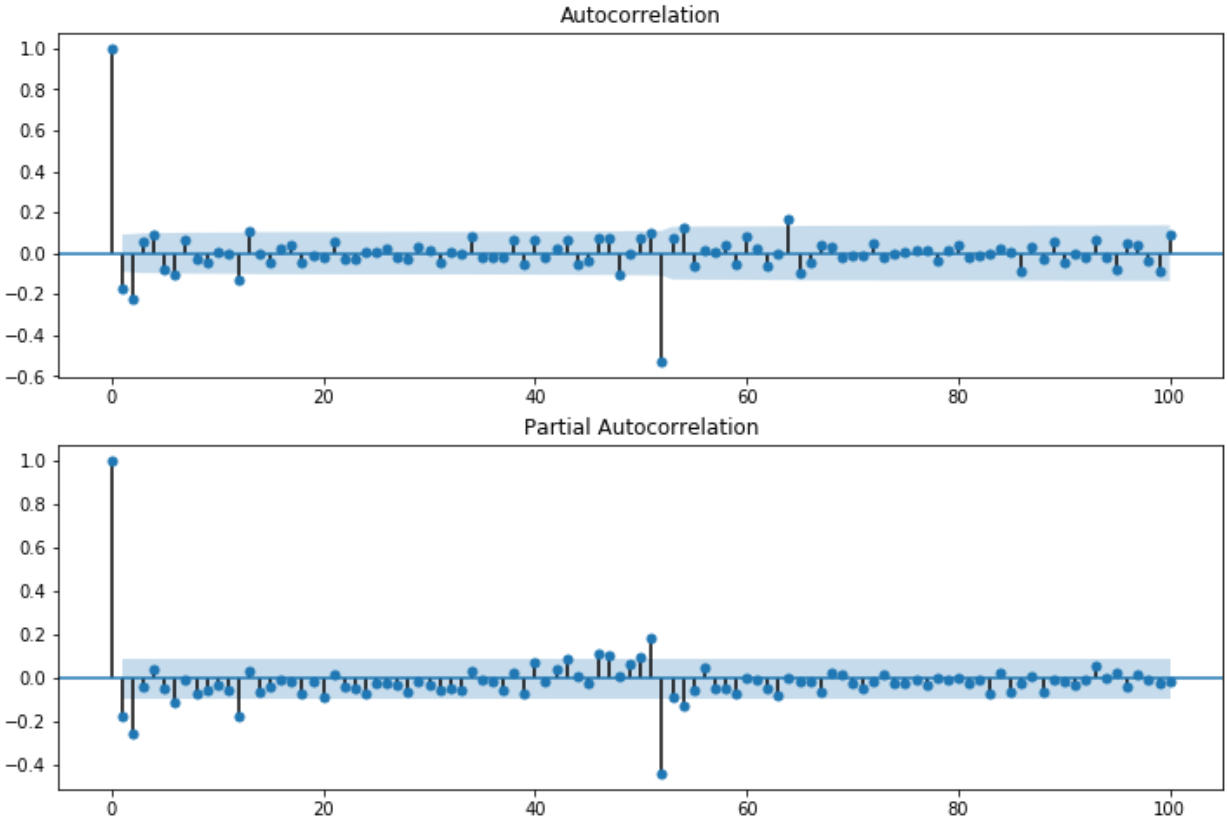


Figure 23. Autocorrelation and Partial Autocorrelation for Iquitos, Peru

When we ran the ARIMA model we got a mean average error of 35. Our worst score out of all the models we used. We quickly realized that the model was not able to forecast accurately more than 20 weeks into the future. Our simple ARIMA model could not account for any sharp changes in rises and falls of cases, and performed especially poorly in Iquitos where there was no strong trend through the years to dengue.

Linear Regression Model

The next model we worked with was the Linear Regression Model. Our general strategy was to try and to look at the monthly spread of dengue, then subtract the monthly trend of dengue cases from the total cases of dengue leaving us with monthly residuals and try and predict the residuals using lagged and rolling features. To do this we looked at the correlations of rolling means and standard deviations for all features in both cities for a span of 52 weeks. We also looked at lagged features to see if there was any correlation there. What we found out was that rolling averages and standard deviations of the features had much stronger correlations to the total cases than did singular value for the features.

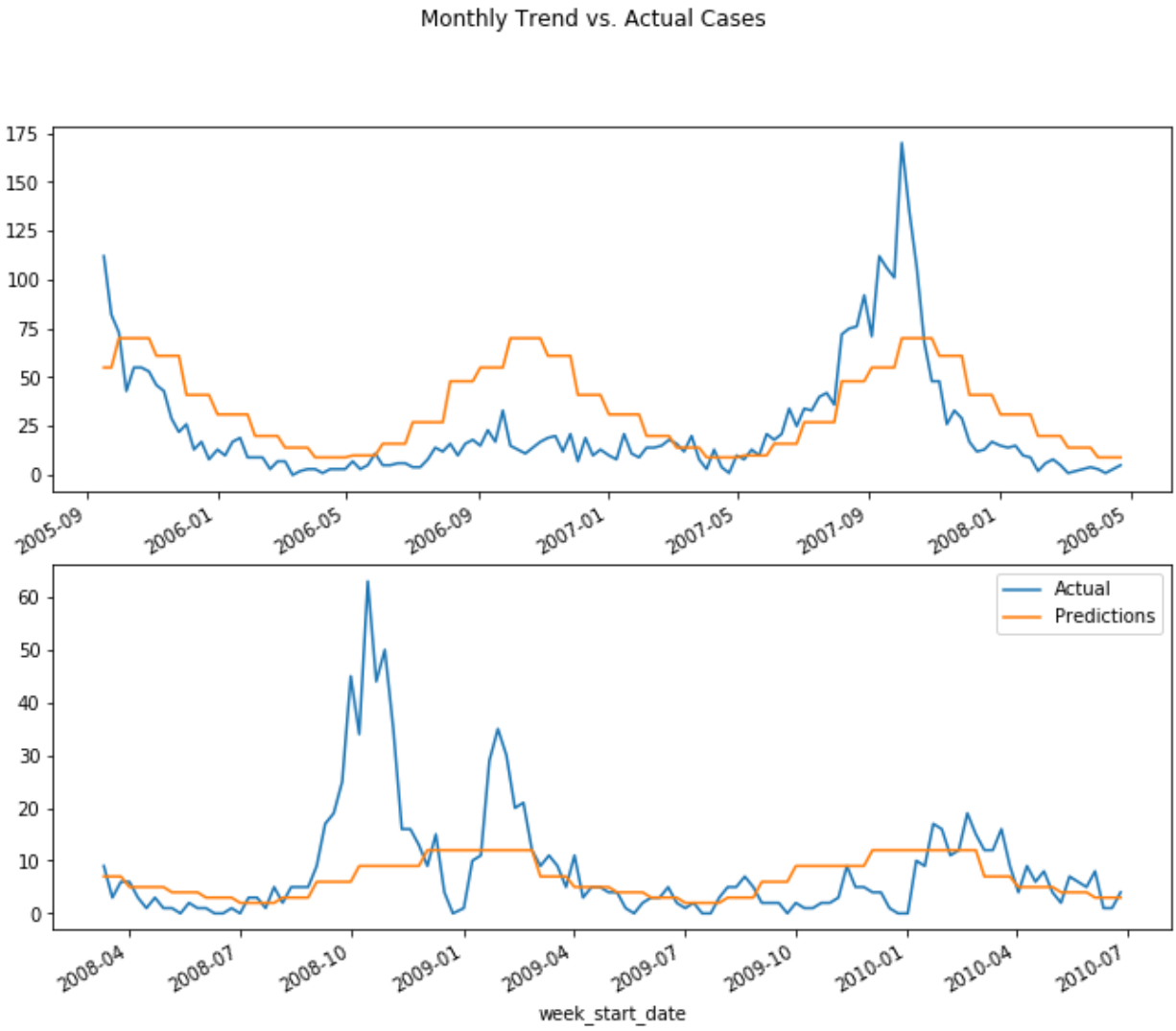


Figure 24. Monthly trend total cases predictions in San Juan (top) and Iquitos (bottom)

As we can see in table 5 and 6, most rolling features were most correlated with total cases as a window of a full year. This was surprising because initially we thought correlated rolling means would be best at smaller window of somewhere between 5-20 weeks. This was only the case for some variables such as the vegetation index for “ndvi_se” which saw its best correlation to total cases when taking the rolling mean of 8 weeks. One theory is that these higher correlations exist when using a larger window because it’s easier to spot a clear stronger positive or linear correlation when the averages are more stable

and have fewer fluctuations as we have with a larger window. Taking a larger window for rolling means impedes our ability to predict short sharp out breaks such in Iquitos, but helps us see a slower developing growth in cases such as exhibited in San Juan.

feature	correlation	window	
12	reanalysis_sat_precip_amt_mm	0.368687	52
4	precipitation_amt_mm	0.368687	52
19	station_precip_mm	0.300305	52
2	ndvi_se	0.271404	8
16	station_diur_temp_rng_c	0.232101	52
17	station_max_temp_c	0.220481	50
15	station_avg_temp_c	0.205247	52

Table 1. Rolling Mean and Window to use for best correlated features for San Juan

feature	correlation	window
2	ndvi_se	0.319816	51
8	reanalysis_max_air_temp_k	0.282091	52
14	reanalysis_tdtr_k	0.27845	52
16	station_diur_temp_rng_c	0.275937	52
11	reanalysis_relative_humidity_percent	0.267481	52
7	reanalysis_dew_point_temp_k	0.260938	52
13	reanalysis_specific_humidity_g_per_kg	0.257218	52
9	reanalysis_min_air_temp_k	0.219504	52
10	reanalysis_precip_amt_kg_per_m2	0.206694	52
17	station_max_temp_c	0.192364	40

Table 2. Rolling Mean and Window to use for best correlated features for Iquitos

We also looked correlations for different windows of standard deviations for all the variables. As we can see in table 7, in San Juan, similarly to rolling means, the strongest correlations exist at the larger

sized windows of a year, however In Iquitos we see a different situation, with the top correlations for standard deviations occurring at a window of between 10 to 40 weeks. Seeing as standard deviation is an indicator of variance and so, in San Juan, variance over a year is more indicative of total cases, while in Iquitos variance from the recent weather patterns (32 weeks for rain, 20 weeks for max temperature) is the best indicator for rises and falls in total cases. One theory is that because cases are relatively low in Iquitos any week with more than 30 cases is considered a spike, meaning that a shorter period of high variance of weather patterns may result in a relatively more noticeable spike in total cases, while in San Juan this shorter window of variance may not produce really noticeable differences.

feature	correlation	window	
4	precipitation_amt_mm	0.332587	52
12	reanalysis_sat_precip_amt_mm	0.332587	52
13	reanalysis_specific_humidity_g_per_kg	0.247682	49
19	station_precip_mm	0.241731	52
7	reanalysis_dew_point_temp_k	0.239095	49
1	ndvi_nw	0.233885	6
5	reanalysis_air_temp_k	0.226294	52
14	reanalysis_tdtr_k	0.215833	52

feature	correlation	window	
6	reanalysis_avg_temp_k	0.215274	52
9	reanalysis_min_air_temp_k	0.207664	52
0	ndvi_ne	0.207542	28
3	ndvi_sw	0.206307	5

Table 3. Rolling Standard Deviation and Window to use for best correlated features for San Juan

16	station_diur_temp_rng_c	0.317382	20
19	station_precip_mm	0.29781	32
17	station_max_temp_c	0.26651	19
4	precipitation_amt_mm	0.23131	52

12	reanalysis_sat_precip_amt_mm	0.23131	52
5	reanalysis_air_temp_k	0.217426	40
2	ndvi_se	0.202204	10
11	reanalysis_relative_humidity_percent	0.19789	41
14	reanalysis_tdtr_k	0.185135	5

Table 4. Rolling Standard Deviation and Window to use for best correlated features for Iquitos

As discussed earlier, multiple previous project studying dengue have recommended taking a look at lagged features to try and predict short outbreaks or seasonal spikes. We followed through on this suggestion and looked at correlations of lagged features and total cases. To our surprise lagged features had lower correlations to total cases than rolling means and rolling standard deviations. We only ended up looking at three lagged features for San Juan, all had to do with vegetation index ndvi_ne, ndvi_nw, and ndvi_se (lagged 36,2, and 39 weeks respectively).

	feature	correlation	window
0	ndvi_ne	0.176878	36
2	ndvi_se	0.151619	2
1	ndvi_nw	0.107411	39

feature	correlation	window

Table 5. Lagged features for San Juan

feature	correlation	window	
0	ndvi_ne	0.176878	36
16	station_diur_temp_rng_c	0.160043	32
2	ndvi_se	0.151619	2
14	reanalysis_tdtr_k	0.111641	51
1	ndvi_nw	0.107411	39

Table 6. lagged features for Iquitos.

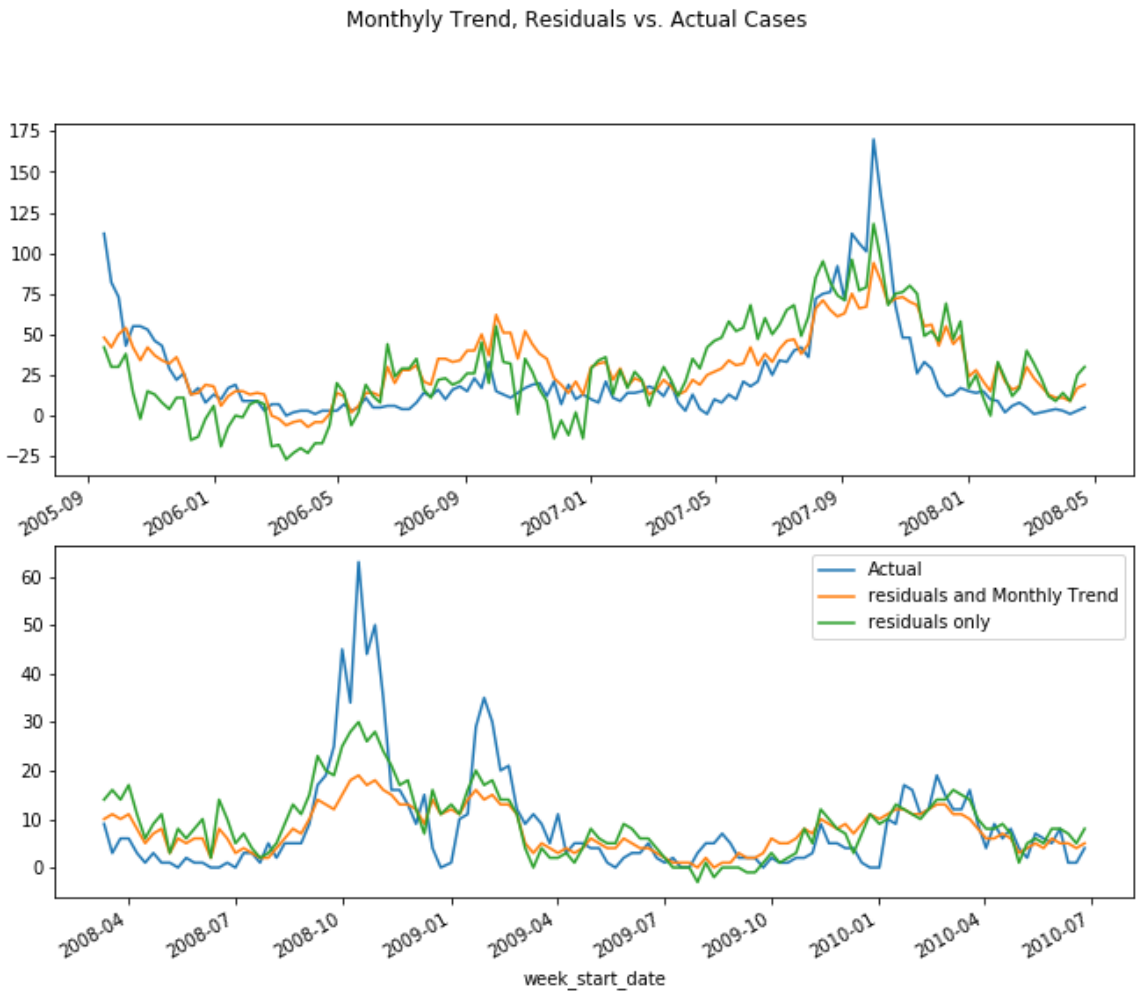


Figure 25. Top: San Juan Predictions for Monthly trend, Residuals and combined

Bottom: Iquitos Predictions for Monthly trend, Residuals and combined

After experimenting with multiple combinations of features, rolling means and standard deviation of features, and lagged features, we created a list of variables to be used for training and for testing specific to each city. We fit the model for monthly trend and predicted on our hold-out test set (as in figure 20) and then did the same for residuals. Out final prediction took the averages of the residual prediction and the monthly trend, as seen in figure 21. This Model yielded us our best MAE of 22.6, and gave us a ranking of 151 out of 2864.

Negative Binomial Regression

The last model we explored was negative binomial regression. A negative binomial regression model is a model in which the distributions parameter is considered a random variable [15]. This makes it so the model will account for cases where the variance in the data is much higher than the mean. When looking at the data both cities showed very large differences in mean compared to variance. As a result, negative binomial regression should be good model for this problem.

DrivenData uses negative binomial regression as their benchmark for this competition. Our implementation was slightly different from theirs, but the results were almost the exact same. We followed their tutorial for this model as a basis for ours.

We start by using the already processed training data from the linear regression model. From here we fill the null values and set up the index. We created a previous week value using the linear regression model predictions. We then take the rolling mean of the last 6 weeks for San Juan and 3 weeks for Iquitos of linearly predicted total_cases. Then we use last week's value to try and predict the current week. We found the features that performed the best, such as reanalysis_max_air_temp_k, station_max_temp_c, reanalysis_min_air_temp_k, reanalysis_air_temp_k and station_min_temp_c among several mean and standard deviations and used those as our model formula for training. We then found the best hyper parameter and alpha and fitted the model.

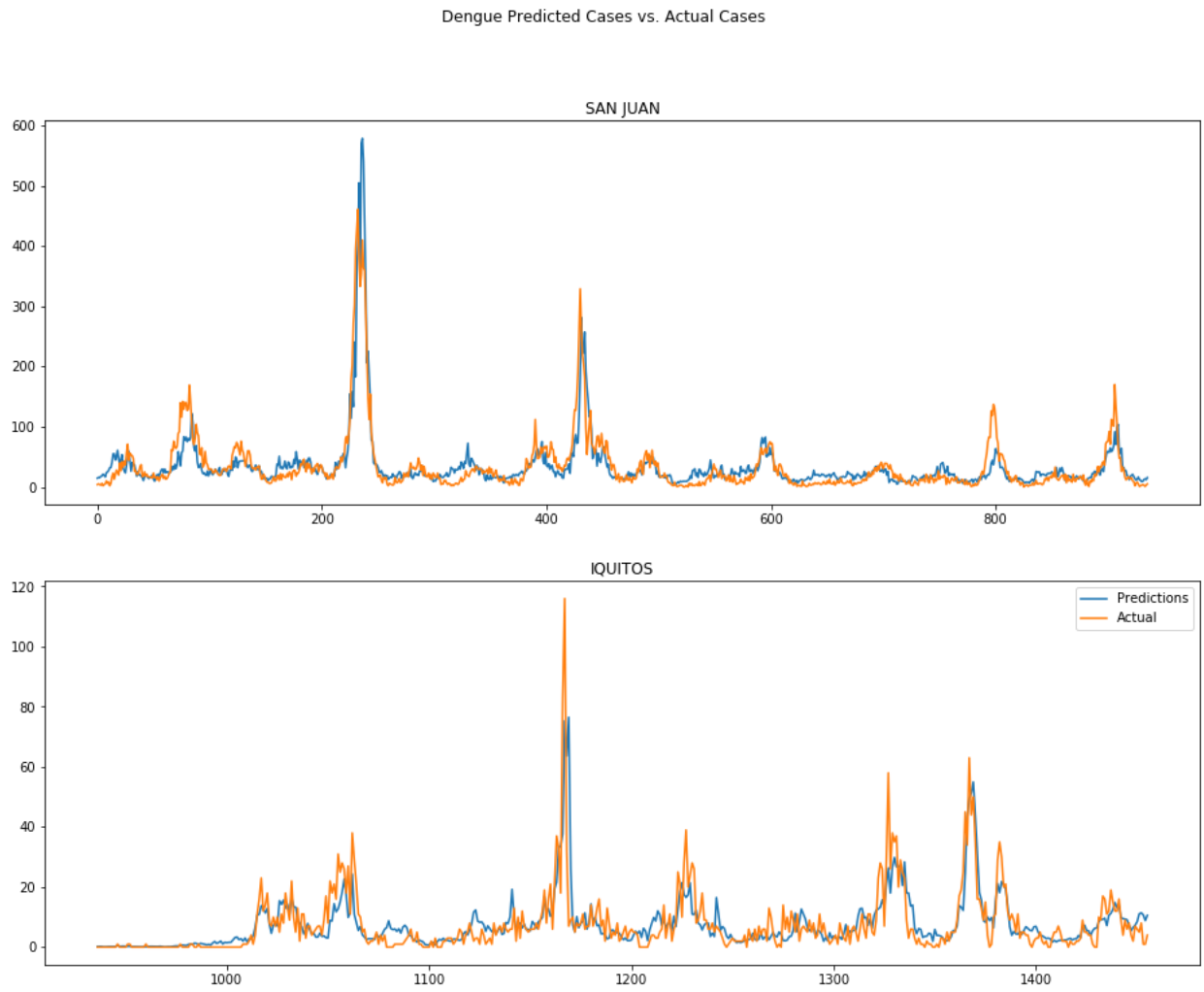


Figure 26. Negative Binomial regression predictions vs actual results

We can see from the results that our model predicts closely to the actual values. The general trend in both cities line up closely to the actual values with peaks that line up in correct time frames. There are some areas where we can see that the peaks, while predicted in the right time frame, are quite different in values. Particularly the biggest spike in both San Juan and Iquitos are either predicted much higher than the actual total_cases, or predicted much lower, respectively. Our score for this model was 24.6 which was better than the DrivenData's result [1]. There are a few reasons for the discrepancies in this model that we will explore in the discussion.

Discussion

We were pleased with the overall results of our models. Our negative binomial regression model performed as expected and in line with the benchmark DrivenData came up with, while our linear regression model outperformed both our expectations as well as both the negative binomial regression model and the ARIMA mode.

A downside of the model supplied by DrivenData is that it misses spikes in total_cases. One possible reason for this is because the model does not address the time lag for mosquitos to transmit the disease, or the time it takes mosquitos to become adults. As discussed in the paper by Karim, Md Nazmul, et al., mosquitos have a 7-45 day period from egg to adult [12]. This time lag means that while more eggs would be laid in the moths with the best conditions for mosquitos to thrive and reproduce in, the results in more actual mosquitos and such higher cases of dengue fever, would not be seen for 1-2 months after. As a result of the inclusion of rolling mean, variance, and lagged features, our negative binomial model does account for these time lags and so it's more likely to predict accurately. While the model did perform and predict well, we were still fairly far off from the actual values. One potential reason for this could be because it does train from data over a long enough period, or at least not over a period far enough away from the predicted weeks. If we had the model look farther in the past for gauging future trends it would have more data to work with and of course strengthen the predictive ability. Another possible downside to both this model and that it doesn't account for monthly trends. With each month of the year having different types of weather, accounting for monthly trends would improve the model's accuracy.

The most successful model we did was the linear regression model. This model used several features to achieve the predictive success and impressive high score. Compared to the benchmark, we looked at the lag time and monthly trends over 52 weeks. The monthly trends were looked at using rolling means and standard deviations. Using rolling means we looked at the weeks leading up the prediction and

averaged the values from those weeks. We then did the same with the standard deviations. By taking a large window for each feature you can get accurate trends over the year, but because of seasonal differences, having smaller windows may give more accurate results for monthly trends for a given feature. By finding strongest correlations for both rolling mean and rolling deviation for each feature in each city, we were able to build a model based on the best window for each rolling feature. The shifting approach mixed with the rolling mean and standard deviation worked very well for predicting total cases, giving us an MAE of 22.6

This is comparable to the paper by Karim, Md Nazmul, et al. []. They used a 1 and 2-month lag window for their comparisons and found that by waiting longer, the results were more accurate. In our case we adjusted for more than just a couple months and so our accuracy is likely going to be better. There would be very little we would have changed with this model to perfect its score. As is, this is the best score we could likely achieve with this type of model.

Figure 27. Linear Regression Score

Our ARIMA model was by far the weakest of the three models. Although the ARIMA model appeared to be a very common and useful model in count data, it did not work out well for us. It was relatively accurate up to 7 weeks, but beyond that the forecast was not good. One reason this model did not perform well is because it only looked at total_cases and not at all of the features. From what our EDA and other models tell us is that there is clearly a strong correlation between the features and total_cases and by just looking at the total_cases we are not getting the whole picture.

To fix this model, we would follow more closely to how the model was done in the paper by Pei-Chih, et al. [10]. They cross-correlated the features for 6 different time-lags and found the ones that were significant. They then concatenated some features, such as maximum temperature and minimum temperature into one variable. Then they made several different models of different covariates to find one for the best fit. Their results with this model were accurate and much better than what we created. We think with some better research and understanding of the ARIMA model we could have created a better functioning model that looked into more than just total_cases.

Conclusion

References

[1] *DengAI: Predicting Disease Spread* [Online]. Available: https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/page/81/

[2] *Using Mean Absolute Error for Forecasting Accuracy* [Online]. Available: http://canworksmart.com/using-mean-absolute-error-forecast-accuracy/

[3] *Dengue Fever* [Online]. Available: https://www.webmd.com/a-to-z-guides/dengue-fever-reference#1

[4] F. Stefan, et al.: "The long-term safety, public health impact, and cost-effectiveness of routine vaccination with a recombinant, live-attenuated dengue vaccine (Dengvaxia): a model comparison study." *PLoS medicine*, Vol 13, No. 11, 2016.

[5] World Health Organization, et al.: "Dengue: guidelines for diagnosis, treatment, prevention and control." *World Health Organization*, 2009.

[6] S. Bhatt, et al.: "The global distribution and burden of dengue." *Nature*, Vol. 496, No. 7446, pp 504, April 2013.

[7] *Back to the Future: Using Historical Dengue Data to Predict the Next Epidemic* [Online]. Available: https://obamawhitehouse.archives.gov/blog/2015/06/05/back-future-using-historical-dengue-data-predict-next-epidemic

[8] *Dengue Forcasting* [Online]. Available: http://dengueforecasting.noaa.gov/

[9] G.P. Zhang: "Time series forecasting using a hybrid ARIMA and neural network model." *Neurocomputing*, Vol. *50*, pp.159-175, 2003.

[10] P.C. Wu, et al.: "Weather as an effective predictor for occurrence of dengue fever in Taiwan." *Acta tropica*, Vol. 103 No.1, pp.50-57, 2007

[11] S. Promprou, M. Jaroensutasinee, and K. Jaroensutasinee: "Climatic Factors Affecting Dengue Haemorrhagic Fever Incidence in Southern Thailand". 2005

[12] M.N. Karim, S.U. Munshi, N. Anwar, and M.S. Alam: "Climatic factors influencing dengue cases in Dhaka city: a model for dengue prediction." *The Indian journal of medical research*, Vol. 136, No.1, pp.32, 2012.

[13] H. Liu, R.A. Davidson, D.V. Rosowsky, and J.R. Stedinger: “Negative binomial regression of electric power outages in hurricanes.” *Journal of infrastructure systems*, Vo. 11, No.4, pp.258-267, 2005

[14] Autoregressive Integrated Moving Average Models (ARIMA) [Online]. Available: http://www.forecastingsolutions.com/arima.html

[15] Regression Models for Count Data [Online]. Available: https://www.theanalysisfactor.com/regression-models-for-count-data/