DengAI: Predicting Disease Spread

Table of Contents

I Problem ***2***

II Significance ***2***

II Data Processing and Missing Values Handling ***2***

IV Descriptive Analysis and Data Visualization ***3***

V Literature ***6***

VI Type of Models and Model Formulation ***6***

VII Performance and Accuracy ***8***

VIII Limitations and Future Work ***9***

IX Learning ***9***

APPENDIX ***9***

References ***11***

I Problem

During the last decades, dengue viruses have spread throughout the sub-tropical parts of the world, with an increase in the number of severe forms of dengue such as severe bleeding, low blood pressure, and even death. Because the disease is carried by mosquitoes, the transmission of Dengue fever is related to meteorological variables such as temperature and precipitation. The surveillance system provided by the U.S. Centers for Disease Control and prevention, and the Department of Defense's Naval Medical Research Unit 6 and the Armed Forces Health Surveillance Center, in collaboration with the Peruvian government and U.S. universities is a good resource for the detection of early outbreaks of dengue. The goal of this research is to utilize the surveillance data by building and assessing models that can be used to forecast the weekly incidents of dengue epidemics in 2 cities: San Juan, Puerto Rico and Iquitos, Peru.

II Significance

As dengue remains a major health risk and outbreaks are increasing, it is urgent to optimize the surveillance in order to control and prevent the occurrence of the disease. An early warning of dengue outbreaks could improve the efficiency of healthcare facilities and help to target prevention actions. An initial step of this strategy is to predict the incidence of dengue cases, thus to help an efficient dengue control. The data set we have is a complicated and messy dataset with 2 different cities, 2 different time frames, and 20 climate variables gathered from 3 different stations, which requires careful examination and analysis. All in all, this is an interesting case as if it is done right, so many lives will be saved, and resources will be allocated more efficiently. To students and researchers, this is a golden chance to work with real life data and help deliver thoughtful insights and solutions on an emerging issue.

II Data Processing and Missing Values Handling

The DengAI competition data set is essentially a messy dataset with two different cities and two different time frames. I split the dataset into two small datasets based on two cities: San Juan and Iquitos. From two datasets, I divided each into train and validation sets in order to test different models before choosing the best one to perform on the final test set.

Each dataset contains 25 variables including meteorological and geographical variables. I handled missing data by using **Last Observation Carried Forward** method in R, which is basically replacing each NA with the most recent non-NA prior to it. Here I am assuming that the weather of the current week would not be too different to the weather of the previous week.

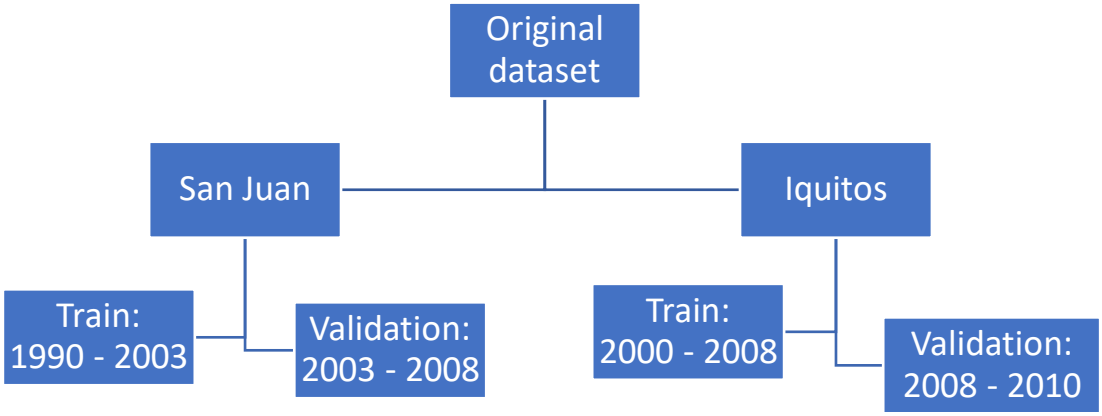


Figure 1: Data Split

IV Descriptive Analysis and Data Visualization

The Data Processing for both San Juan and Iquitos were quite similar, so I will only show San Juan in this section. Charts and Graphs of Iquitos will be provided in the Appendix.

San Juan

A Correlation plot will be helpful to diagnose if there are any linear relationships between 2 variables. We can see none of those variables are highly correlated to total cases **(Figure 3)**. However, if we include more than 2 variables and run multivariate linear regression, the magnitude of each predictors might change.

The time series plot indicates that there might be seasonality in the data since the number of total cases peaks every year around summer time. Two scatter plots also suggest possible linear relationships between total cases and two climate variables.

Looking at the time series plots of total cases and temperature, and total cases and humidity, we can see sometimes the number of cases and climate values move up and down together; sometimes the weather variables peak or plummet first, then the total cases go up and down in 3 -4 weeks later. More interestingly, Dengue symptoms usually start 4 to 7 days after you are bitten by an infected mosquito. Mosquito might go through its life cycle in 14 days at 70° F and take only 10 days at 80° F. Therefore, in average, it takes 3-4 weeks for the cases to be reported after the temperature, precipitation or humidity go up and down.

The Descriptive Analysis and Data Visualization suggest a time series model with predictor variables to included, and possibly some previous lags of the predictors.

San Juan - Descriptive statistics

Statistic	N	Mean	St. Dev.	Min	Max
year	659	1,996.2	3.7	1,990	2,002
weekofyear	659	27.0	14.9	1	53
ndvi_ne	520	0.1	0.1	-0.3	0.4
ndvi_nw	620	0.1	0.1	-0.3	0.4
ndvi_se	641	0.2	0.1	0.04	0.4
ndvi_sw	641	0.2	0.1	-0.1	0.4
precipitation_amt_mm	652	34.2	41.1	0.0	287.6
reanalysis_air_temp_k	655	299.0	1.2	295.9	301.3
reanalysis_avg_temp_k	655	299.1	1.2	296.1	301.4
reanalysis_dew_point_temp_k	655	295.1	1.6	289.6	297.5
reanalysis_max_air_temp_k	655	301.3	1.2	298.2	303.9
reanalysis_min_air_temp_k	655	297.2	1.3	292.6	299.5
reanalysis_precip_amt_kg_per_m2	655	32.0	37.2	0.0	570.5
reanalysis_relative_humidity_percent	655	79.0	3.4	66.7	87.3
reanalysis_sat_precip_amt_mm	652	34.2	41.1	0.0	287.6
reanalysis_specific_humidity_g_per_kg	655	16.5	1.5	11.7	19.0
reanalysis_tdtr_k	655	2.4	0.5	1.4	4.1
station_avg_temp_c	655	27.1	1.4	22.8	30.1
station_diur_temp_rng_c	655	6.9	0.8	4.5	9.9
station_max_temp_c	655	31.7	1.6	27.2	35.6
station_min_temp_c	655	22.6	1.5	17.8	25.6
station_precip_mm	655	25.1	27.5	0.0	305.9
total_cases	659	39.5	57.9	0	461

Figure 2: Descriptive Analysis of San Juan

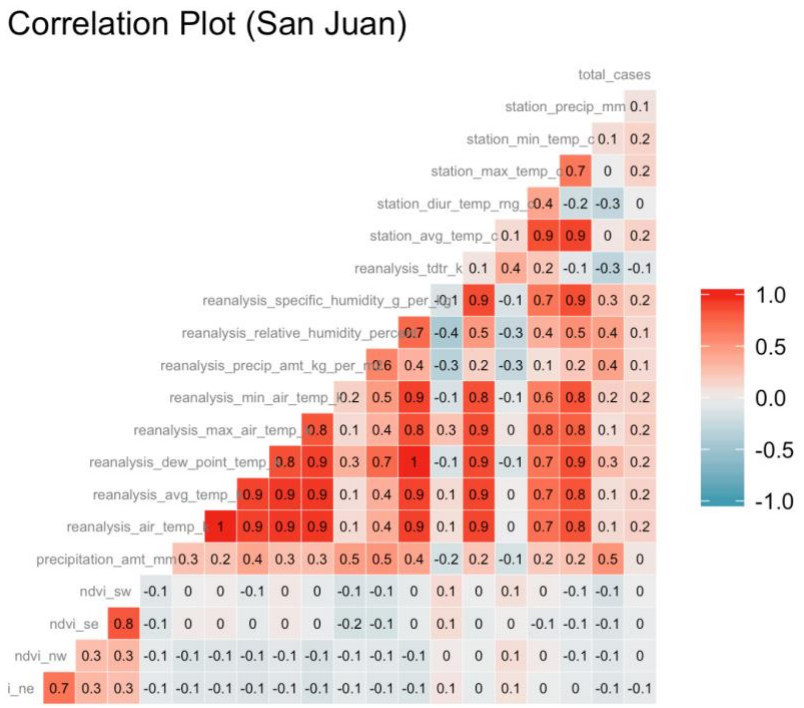


Figure 3: Correlation Plot of San Juan

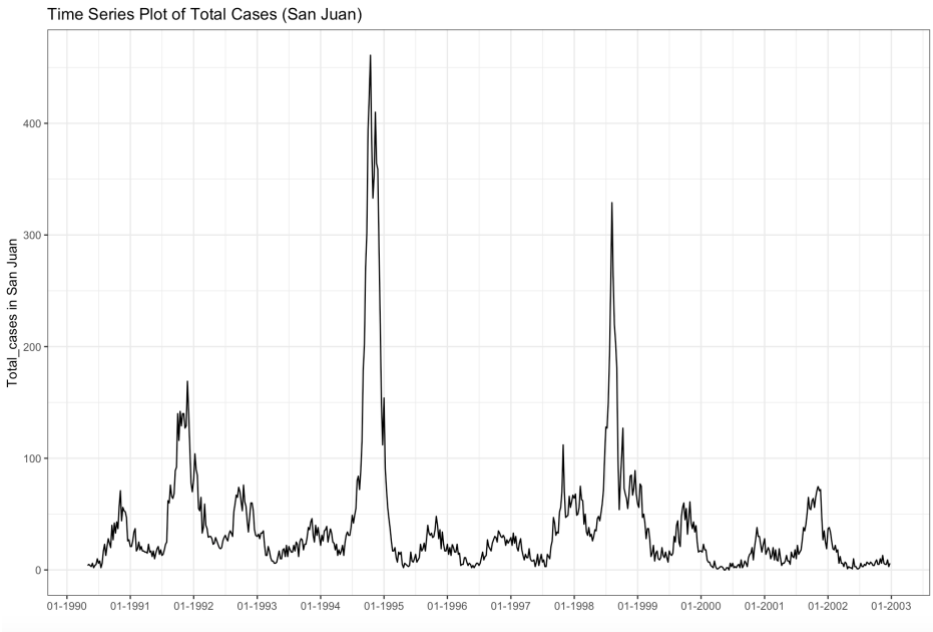


Figure 4: Time Series Plot of Total Cases of San Juan

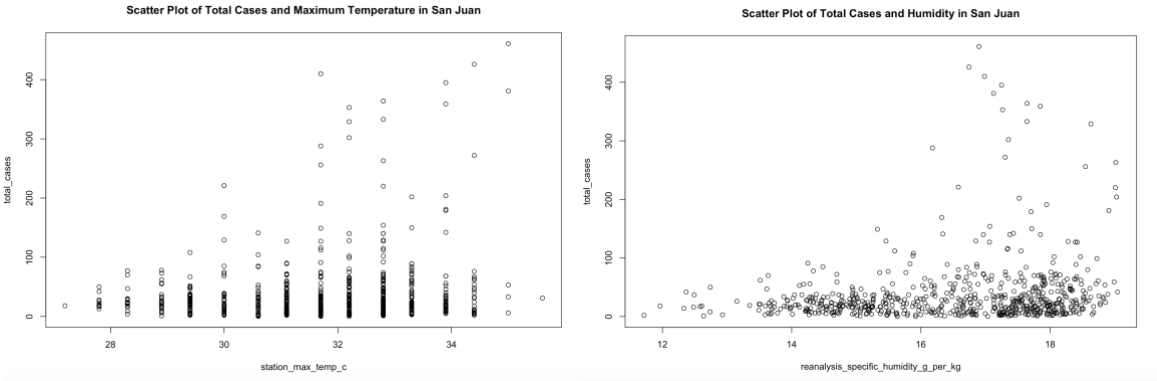


Figure 5: Scatter Plots of Total Cases and Climate Variables of San Juan

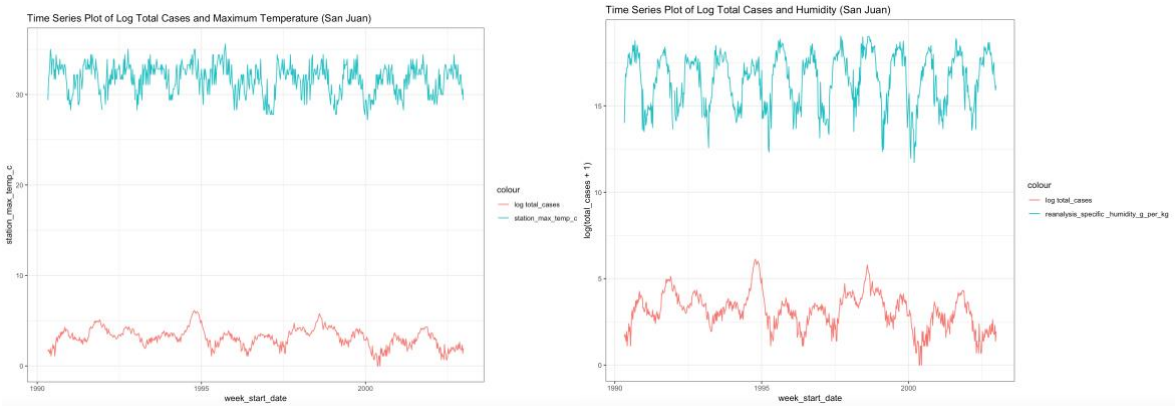


Figure 6: Time Series Plot of Total Cases and Climate Variables of San Juan

V Literature

Due to the complex seasonality of disease spread data, seasonal ARIMA (SARIMA) model has been extensively used in academic researches as well as real world surveillance systems. Many studies discover the relationship between some disease such as Dengue, Hand-Foot-Mouth, and Cholera using Multivariate ARIMA (ARIMAX). Researchers have been avoiding using ETS or Decomposition model for this particular problem since those models do not handle seasonal data with large frequency (larger than 12) very well. In the recent years, with the improvement and spread of Statistical and Machine Learning Packages, some researchers have presented Artificial Neural Networks models to predict disease risk. Artificial neural networks are forecasting methods that are based on mathematical models of the brain, which is helpful when the relationships between the response variable and its predictors are nonlinear. Here are 6 examples of all the models I mentioned above:

Zhang X, Liu Y, Yang M, Zhang T, Young AA, Li X (2013) Comparative Study of Four Time Series Methods in Forecasting Typhoid Fever Incidence in China. PLoS ONE 8(5): e63116. https://doi.org/10.1371/journal.pone.0063116

Pezeshki, Zahra; Tafazzoli-Shadpour, Mohammad; Nejadgholi, Isar; Mansourian, A LU and Rahbar, Mohammad (2016) In International Journal of Enteric Pathog 4(1). p.23-30

Gharbi M, Quenel P, Gustave J, et al. Time series analysis of dengue incidence in Guadeloupe, French West Indies: Forecasting models using climate variables as predictors. BMC Infectious Diseases. 2011;11:166. doi:10.1186/1471-2334-11-166.

Feng H, Duan G, Zhang R, Zhang W (2014) Time Series Analysis of Hand-Foot-Mouth Disease Hospitalization in Zhengzhou: Establishment of Forecasting Models Using Climate Variables as Predictors. PLoS ONE 9(1): e87916. https://doi.org/10.1371/journal.pone.00

Wenbiao Hu, Shilu Tong, Kerrie Mengersen, Des Connell, Weather Variability and the Incidence of Cryptosporidiosis: Comparison of Time Series Poisson Regression and SARIMA Models, Annals of Epidemiology, Volume 17, Issue 9, 2007, Pages 679-688, ISSN 1047-2

Sudarat Chadsuthi, Charin Modchang, Yongwimon Lenbury, Sopon Iamsirithaworn, Wannapong Triampo, Modeling seasonal leptospirosis transmission and its association with rainfall and temperature in Thailand using time–series and ARIMAX analyses, Asian Pacific

VI Type of Models and Model Formulation

With the empirical review and data diagnostics, I decided to try 3 models: SARIMA, SARIMAX, and Neural Networks. I started with a basic and simple Seasonal ARIMA with only total cases, then I added more predictors such as temperature, humidity, precipitation into the model. Lastly, I built Neural Networks model (multilayer feed-forward network) with NNETAR package in R.

SARIMA - ARIMA(p, d, q) × (P, D, Q)S

A seasonal ARIMA model was formed by including additional seasonal terms in the ARIMA models. As the ACF and PACF suggest that there is seasonality effect in the data, I include “seasonal = T” in auto.arima. The seasonal part of the ARIMA model consists of terms that are similar to the non-seasonal components of the model but involve backshifts of the seasonal period. I also ran a cross validation code to pick the best Fourier term to account for complex seasonal effect in the dataset.

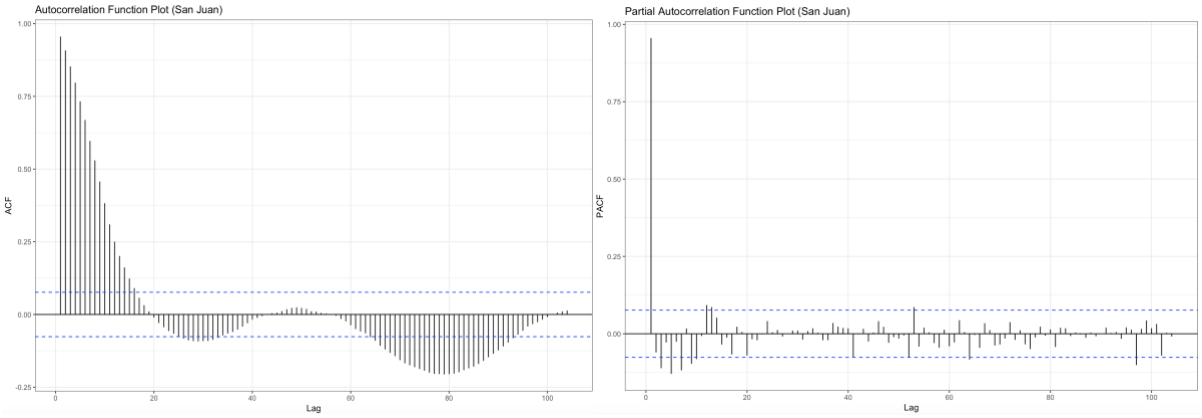


Figure 7: ACF and PACF Plots of Total Cases of San Juan Time Series

SARIMAX

$$y_t = \beta_0 + \gamma_0 x_t + \gamma_1 x_{t-1} + \cdots + \gamma_k x_{t-k} + \eta_t$$

Here after doing researches about what factors affect the proliferation of mosquitos, I added on these variables to SARIMA using xreg arguments:

- precipitation_amt_mm
- reanalysis_dew_point_temp_k
- reanalysis_relative_humidity_percent
- station_avg_temp_c
- station_diur_temp_rng_c
- station_max_temp_c
- station_min_temp_c

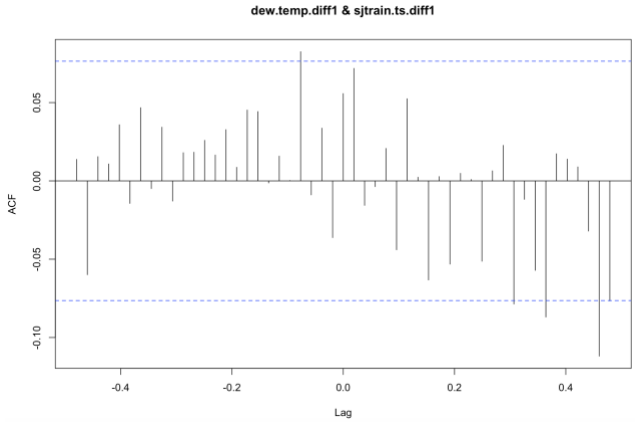


Figure 8: CCF Plots of Total Cases and Dew Temperature of San Juan Time Series

Sometimes, the impact of a predictor which is included in a regression model will not be simple and immediate. For example, the high temperature and high quantity of rainfall may impact total cases for some time after a few weeks. I run CCF to see if any previous lags of the predictors are highly correlated with the current lag of total cases.

Neural Networks - multilayer feed-forward network
Two Neural Networks Models were conducted for 2 cities using all numeric variables. The function nnetar builds a network of "neurons" which are organized in layers. The predictors form the bottom layer, and the forecasts form the top layer. There may also be intermediate layers containing "hidden layers".

VII Performance and Accuracy

The selection of ARIMA models was conducted using Akaike's information criterion (AIC). Of all the model tested, a SARIMAX (3,0,2)(0,0,1)[52] are the most appropriate model for San Juan city, while ARIMAX (0,1,2) with K = 1 for seasonal Fourier term are the best model for Iquitos city. The final choice between ARIMA and NNET based on in sample cross validation using MAE as the main parameter.

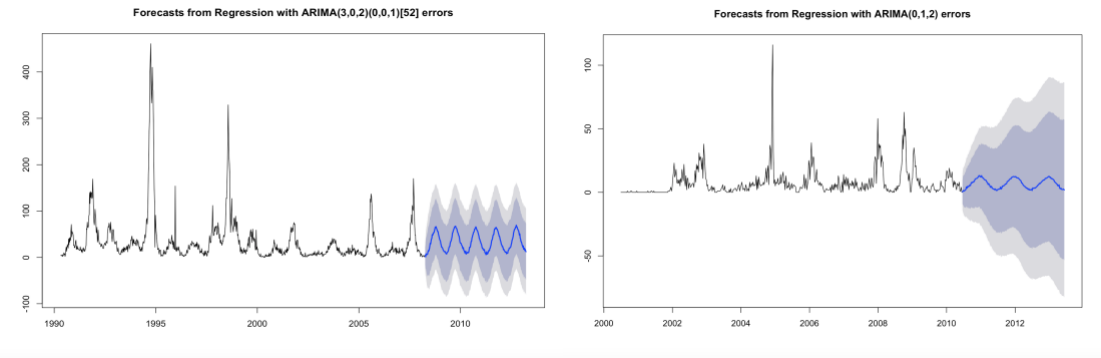


Figure 9: Forecasted Valus of Total Cases of San Juan

In sample

AICc	Model without climate variables	Model with climate variables		Best Model
	SARIMA	SARIMAX	NNET	
San Juan	5589.33	5571.82	x	SARIMAX
Iquitos	2624.9	2638.18	x	SARIMA

Out of sample forecast

MAE	Models			Best Model
	SARIMA	SARIMAX	NNET	
San Juan	35.446223	21.741999	42.065767	SARIMAX
Iquitos	7.821991	8.130527	10.2540033	SARIMA

Driven Data Submission (MAE)

SARIMA (IQ) + SARIMAX (SJ)	SARIMAX (Both cities)	NNET (Both cities)
25.6587	25.6587	32.1538

VIII Limitations and Future Work

The final model has some shortcomings that can be addressed in the future. Firstly, I totally left out “ndvi” (geographical variables and Satellite vegetation index) while these variables might be important to the breakouts of the disease. However, to include those variables, I need to do more researches to have a better understanding of what, why, and how it matters. Secondly, the model does not fully capture all the factors contributing to the spread of Dengue. More variables related to the infrastructure and healthcare systems of both cities should be included in the model. Lastly, a multiple seasonal period model could be conducted to show different seasonal pattern of week, month and year as the current model only captures the weekly seasonality.

IX Learning

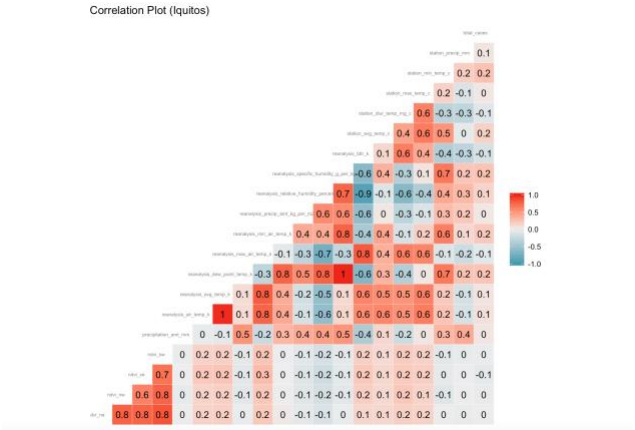
The most valuable I learnt from this competition is that building a robust model requires not only statistical and econometrics knowledge but also an in-depth understanding of the problem. For example, to be able to pick the appropriate predictors, we need to know what factors cause the dengue incidence and accelerate the disease dissemination.

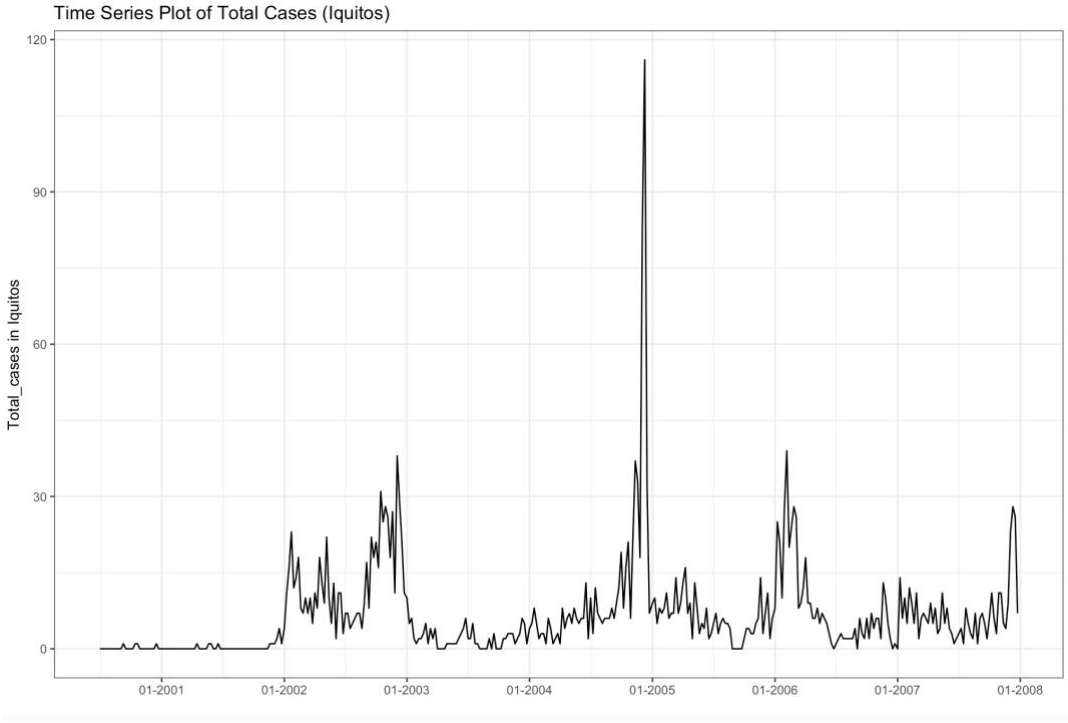
APPENDIX

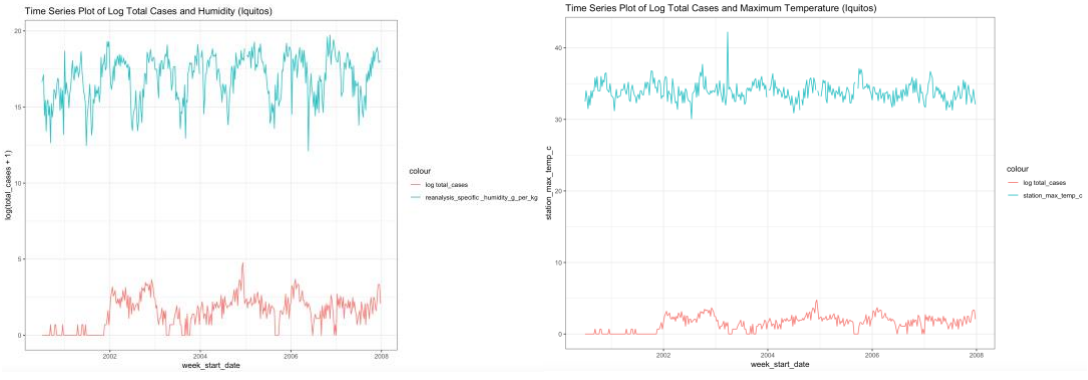
Iquitos’ Graphs and Plots

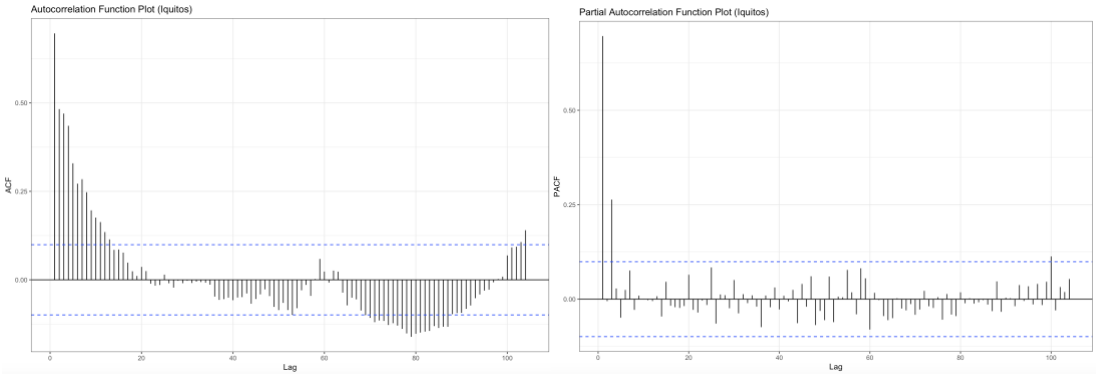
Iquitos – Descriptive Statistics

Statistic	N	Mean	St. Dev.	Min	Max
year	390	2,003.7	2.2	2,000	2,007
weekofyear	390	27.3	15.0	1	53
ndvi_ne	388	0.3	0.1	0.1	0.5
ndvi_nw	388	0.2	0.1	0.04	0.4
ndvi_se	388	0.3	0.1	0.03	0.5
ndvi_sw	388	0.3	0.1	0.1	0.5
precipitation_amt_mm	388	64.0	35.1	0.0	173.4
reanalysis_air_temp_k	388	297.9	1.2	294.6	301.6
reanalysis_avg_temp_k	388	299.2	1.3	294.9	302.9
reanalysis_dew_point_temp_k	388	295.3	1.5	290.1	297.9
reanalysis_max_air_temp_k	388	307.2	2.4	300.0	314.0
reanalysis_min_air_temp_k	388	292.7	1.7	286.9	295.7
reanalysis_precip_amt_kg_per_m2	388	52.8	49.7	0.0	362.0
reanalysis_relative_humidity_percent	388	87.8	8.0	57.8	98.5
reanalysis_specific_humidity_g_per_kg	388	16.9	1.5	12.1	19.7
reanalysis_tdtr_k	388	9.5	2.5	4.2	16.0
station_avg_temp_c	364	27.5	0.9	21.4	30.8
station_diur_temp_rng_c	364	10.5	1.5	5.2	15.8
station_max_temp_c	384	33.9	1.3	30.1	42.2
station_min_temp_c	384	21.2	1.4	14.7	24.2
station_precip_mm	381	69.0	67.4	0.0	543.3
total_cases	390	6.6	9.9	0	116


Correlation Plot (Iquitos)


Time Series Plot of Total Cases (Iquitos)
Total_cases in Iquitos
120
90
60
30
0
01-2001
01-2002
01-2003
01-2004
01-2005
01-2006
01-2007
01-2008


Time Series Plot of Log Total Cases and Humidity (Iquitos)
log(total_cases + 1)
week_start_date
colour
log total_cases
reanalysis_specific_humidity_g_per_kg
Time Series Plot of Log Total Cases and Maximum Temperature (Iquitos)
station_max_temp_c
week_start_date
colour
log total_cases
station_max_temp_c



References

Zhang X, Liu Y, Yang M, Zhang T, Young AA, Li X (2013) Comparative Study of Four Time Series Methods in Forecasting Typhoid Fever Incidence in China. PLoS ONE 8(5): e63116. https://doi.org/10.1371/journal.pone.0063116

Pezeshki, Zahra; Tafazzoli-Shadpour, Mohammad; Nejadgholi, Isar; Mansourian, A LU and Rahbar, Mohammad (2016) In International Journal of Enteric Pathog 4(1). p.23-30

Gharbi M, Quenel P, Gustave J, et al. Time series analysis of dengue incidence in Guadeloupe, French West Indies: Forecasting models using climate variables as predictors. BMC Infectious Diseases. 2011;11:166. doi:10.1186/1471-2334-11-166.

Feng H, Duan G, Zhang R, Zhang W (2014) Time Series Analysis of Hand-Foot-Mouth Disease Hospitalization in Zhengzhou: Establishment of Forecasting Models Using Climate Variables as Predictors. PLoS ONE 9(1): e87916. https://doi.org/10.1371/journal.pone.00

Wenbiao Hu, Shilu Tong, Kerrie Mengersen, Des Connell, Weather Variability and the Incidence of Cryptosporidiosis: Comparison of Time Series Poisson Regression and SARIMA Models, Annals of Epidemiology, Volume 17, Issue 9, 2007, Pages 679-688, ISSN 1047-2

Sudarat Chadsuthi, Charin Modchang, Yongwimon Lenbury, Sopon Iamsirithaworn, Wannapong Triampo, Modeling seasonal leptospirosis transmission and its association with rainfall and temperature in Thailand using time–series and ARIMAX analyses, Asian Pacific