Predict 413 Sect. 55:

Predicting the Spread of Dengue Fever in San Juan, Puerto Rico & Iquitos, Peru

Daren Purnell

Northwestern University

Executive Summary & Introduction

The purpose of this analysis is to predict the number of cases of Dengue Fever in San Juan, Puerto Rico and Iquitos, Peru given environmental variables for a specific week of the year. Because the transmission method of Dengue Fever is mosquitos, the transmission dynamics are related to environmental variables such as vegetation, precipitation, humidity, and minimum & maximum temperatures. In this paper, we will review the exploratory data analysis, data preparation, model building, and model selection processes that led us develop and build our predictive model.

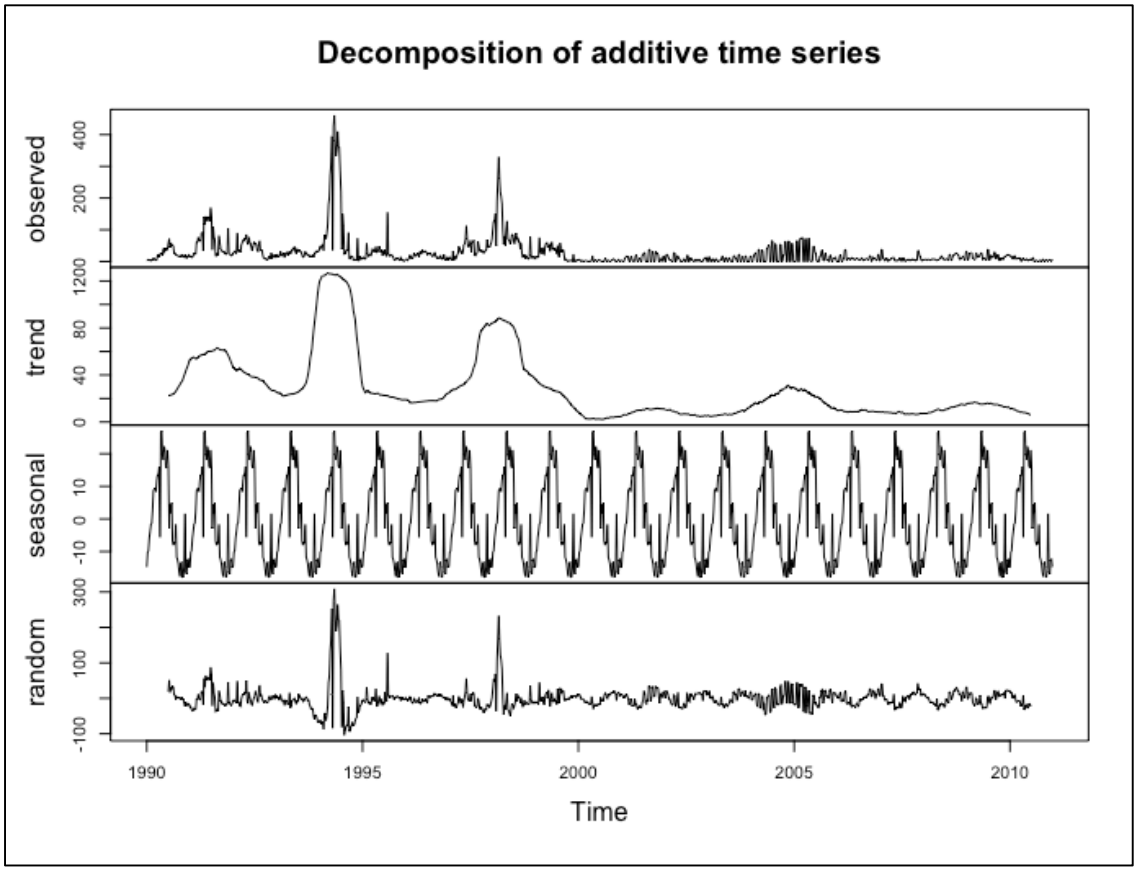
Dengue Fever is a mosquito-borne illness that originates in tropical and subtropical climates. Contracting the virus can result in fever, rash, aches, and in the most severe cases death. In the past, Dengue was most prevalent to Southeast Asia and Pacific Islands however, it has recently spread to Latin America with over a half-billion cases per year being reported. Because the transmission rate of the illness is related to the mosquito population, many researchers believe there is a link between the spread of the virus and climate change. By developing a better understanding of the environmental factors that contribute to the spread of Dengue Fever, Governments, Non-Governmental Institutions, and Researchers can develop policy and procedures to prevent the next Dengue Fever pandemic.

Exploratory Data Analysis

The Dengue Fever data set consist of the response variable, total_cases, and 25 other potential regressor variables that indicate the city of occurrence (San Juan or Iquitos) and several environmental factors relating to vegetation growth, humidity, and temperature that span the period from 1990 to 2010. There is a total of 2270 records and several of the regressors are missing values.

A plot of the regressor variable, total_cases, time series components shows potential pandemics of Dengue occurring roughly around the early 1990s, 1994, and 1997. The period following the potential outbreak in 1997 to 2010 has been relatively stable with a moderate increase of the number of cases of Dengue occurring in 2005. There is a clear seasonal cycle of total cases that peaks during the summer months, with warmer weather, and declines as the weather cools. This seasonal cycle adheres to our assumption that warmer, wetter, weather contributes to more mosquitos and cases of Dengue among the resident population. The trend of total_cases does not have a consistent slope (rate of Dengue spread), rather there are several changes of directions that appear to be consistent with past outbreaks of Dengue Fever. There may be several reasons for this change in trend ranging from implementation of specific policies to address the spread of mosquitos and Dengue Fever to simple fluctuations in weather patterns that created pools of water for mosquitos to breed and reproduce. The random, or noise, portion of the decomposition appears to track with observed plot of total_cases which indicates that there is not a great deal of random noise in our total_cases time series.

Decomposition of Total_Case Time Series



Computing the Pearson Correlation Coefficients of our data set allow us to better understand the linear relationship between the response and regressor variables. All the vegetation index variables (ndvi_ne, ndvi_nw, ndvi_se, ndvi_sw) appear to have a minimal inverse relationship with response variable total_cases. This minimal relationship will lead us to remove this variables from our analysis as they appear to not significantly contribute to the number of cases of Dengue Fever. Counter to intuition, the precipitation related regressors (precipitation_amt_mm, reanalysis_precip_amt_kg_per_m2, reanalysis_sat_precip_amt_mm, station_precip_mm) have a minimal inverse relationship with total_cases. This seems strange because large pools of water provide places for mosquitos to lay their eggs and reproduce, thus increasing the number of Dengue transmitters. We will also remove these variables from our initial analysis because they appear to have minimal association with our response variable total_cases. The regressor variables that relate to air temperature (reanalysis_air_temp_k, reanalysis_avg_temp_k, reanalysis_max_air_temp_k, reanalysis_min_air_temp_k, station_avg_temp_c, station_diur_temp_rng_c, station_max_temp_c , station_min_temp_c, reanalysis_tdtr_k) with an emphasis on the minimum temperatures had a relatively strong linear relationship with the response variable. I'm hypothesizing that below a certain minimum temperature, the mosquito eggs and larvae die thereby reducing the overall population of Dengue Fever transmitters. The predictor variables that relate to humidity (reanalysis_relative_humidity_percent, reanalysis_specific_humidity_g_per_kg, reanalysis_relative_humidity_percent, reanalysis_air_temp_k

, reanalysis_dew_point_temp_k) also appear to have relatively strong relationship with the response variable total_cases. The computation and analysis of the Pearson Correlation Coefficients, has led me to believe that warmer, more humid temperatures contribute to the spread of Dengue Fever in San Juan, Puerto Rico and Iquitos, Peru.

Regressor Variable R-value (Linear Relationship) with Response Variable Total_Cases

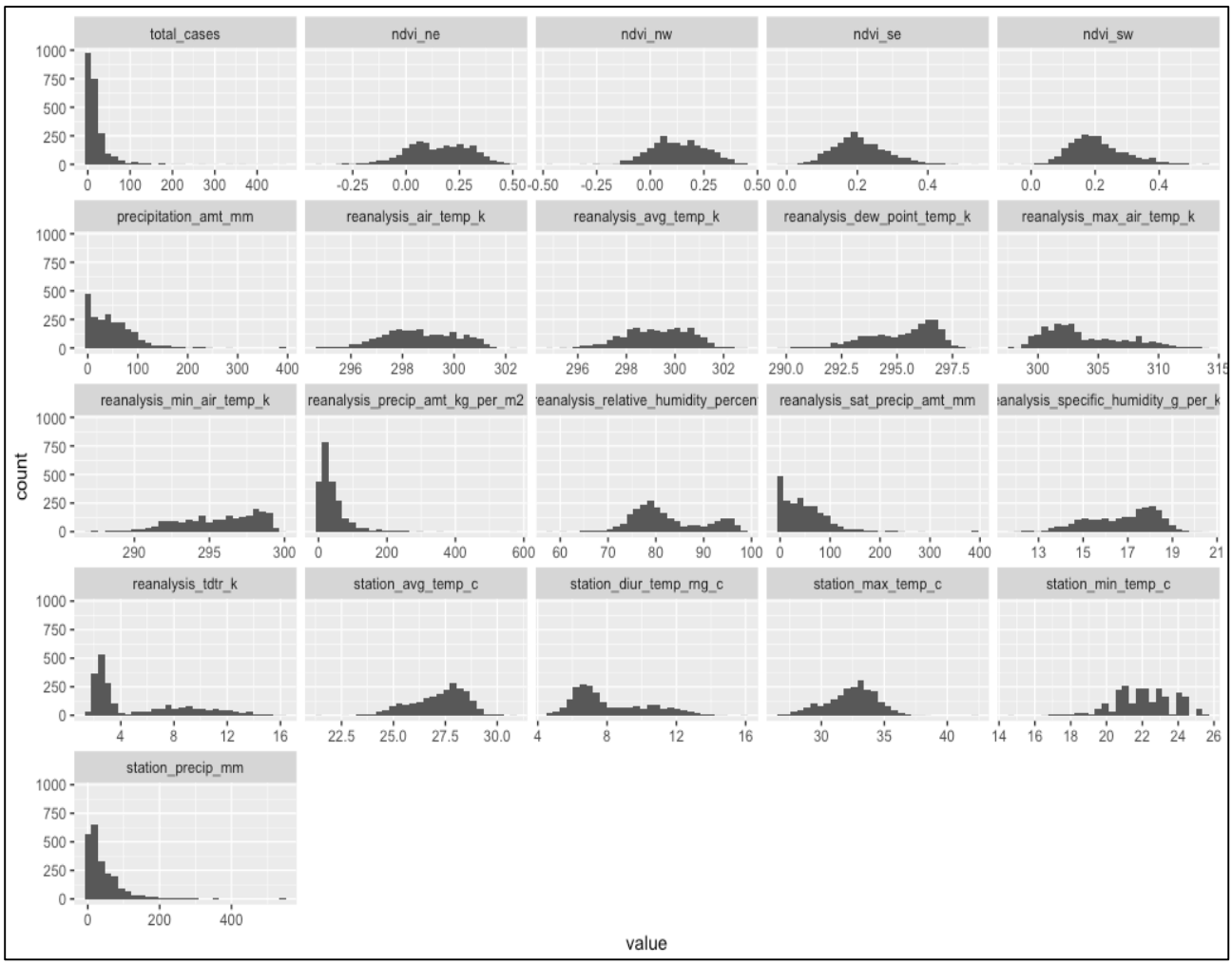
Regressor	Description	R-Value	Interpretation
ndvi_ne	Vegetation Index	-0.046714997	Minimum inverse relationship
ndvi_nw	Vegetation Index	-0.042877295	Minimum inverse relationship
ndvi_se	Vegetation Index	-0.043212145	Minimum inverse relationship
ndvi_sw	Vegetation Index	-0.039612631	Minimum inverse relationship
precipitation_amt_mm	Total Precipitation	-0.041827877	Minimum inverse relationship
reanalysis_air_temp_k	Mean dew point temperature	0.177802573	Positive relationship
reanalysis_avg_temp_k	Average Air Temp	0.118924127	Positive relationship
reanalysis_dew_point_temp_k	Mean dew point temperature	0.074391221	Minimum positive relationship
reanalysis_max_air_temp_k	Maximum air temperature	-0.082905557	Minimum inverse relationship
reanalysis_min_air_temp_k	Minimum air temperature	0.188295854	Positive relationship
reanalysis_precip_amt_kg_per_m2	Total precipitation	-0.00599653	Minimum inverse relationship
reanalysis_relative_humidity_percent	Mean relative humidity	-0.105465066	Inverse relationship
reanalysis_sat_precip_amt_mm	Total precipitation	-0.041827877	Minimum inverse relationship
reanalysis_specific_humidity_g_per_kg	Mean specific humidity	0.067236996	Minimum positive relationship
reanalysis_tdtr_k	Diurnal temperature range	-0.141058381	Inverse relationship
station_avg_temp_c	Weather Station Average Temperature	0.085594348	Minimum positive relationship
station_diur_temp_rng_c	Weather Station Diurnal Temperature	-0.112020413	Inverse relationship
station_max_temp_c	Weather Station Maximum Temperature	0.003662314	Minimum positive relationship
station_min_temp_c	Weather Station Minimum Temperature	0.147387888	Positive relationship
station_precip_mm	Weather Station Total Precipitation	-0.048949078	Minimum inverse relationship

Data Preparation

Data preparation efforts consisted of variable transformation for ordinary least squares regression and imputation of missing records. While not necessary for generalized linear models, Box Cox variable transformation produced an improved predictive model over other methods of transformation. Logistic, additive, and multiplicative transformation were attempted, however the Box Cox transformation proved to be superior. Adhering to the assumptions of linear regression (i.e. normal and independent distribution, linear relationship between variables, no or limited autocorrelation/multicollinearity, homoscedasticity) produced a more accurate model. Rather than using measure of central tendency to replace missing records, I choose to use the last value in the regressors' time series to replace any missing records. This imputation methodology seemed the most logical because it reflected the tendencies of that variable, at that

specific week in time. No outliers were removed from the data to their effects upon the selected Linear, Negative-Binomial, and Poisson Regression models.

Regressor Variable Distributions



Model Development & Selection

For our analysis, we built multiple multivariate linear regression, Neural Network, and time series models utilizing Poisson, Negative Binomial regression and ARIMA with both untransformed and transformed (box cox transformation) variables. Following variable transformation and imputation, we progressed to splitting the designated training data 70/30 with 70% composing the revise training set and 30% creating a new validation data set. This validated data was used to compare each model's predictions against the validated data's total_cases, using the mean absolute error (MAE) to assess goodness of fit. Due to the number of variables in the provided data set, we selected our chosen model's regressor variables based off the strength of their linear relationship with the response variable total_cases. A summary of the accuracy outcomes is below for comparison:

Model Type	MAE
General Linear Regression	15.53116
Negative Binomial	12.4032
Poisson	12.40738
Neural Network w/ Log Transformations	14.09859
Gen. Linear Regression w/ Box Cox Transformations	14.53373
Negative Binomial w/ Box Cox Transformations	12.46298
Poisson w/ Box Cox Transformation	12.48098
Neural Network	15.14978
ARIMA of Total_Cases Time-Series (2,1,3)	32.35923
Dynamic Regression ARIMA	21.26608
Neural Network w/ Log & Box Cox Transformations	19.40339
Negative Binomial w/ Box Cox Transformation of reanalysis_min_air_temp	12.402

Our selected model, is a multivariate Negative Binomial linear regression model incorporating a box-cox transformation of the regressor variable reanalysis_min_air_temp that takes the form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \varepsilon$$

In Model	In Data	Beta	Value
Y	Total_Cases	β_0	-7.936e+01
X1	reanalysis_air_temp_k	β_1	5.510e-01
X2	reanalysis_avg_temp_k	β_2	-2.693e-01
X3	reanalysis_min_air_temp_k_BC	β_3	-1.008e-77
X4	reanalysis_tdtr_k	β_4	-4.913e-02
X5	station_diur_temp_rng_c	β_5	-4.678e-03
X6	station_min_temp_c	β_6	-4.372e-02
ε	Error Term		

Model Notes:

- reanalysis_avg_temp_k has a negative coefficient which is counter to intuition. I would hypothesize that warmer temperatures would contribute to more cases of Dengue by creating favorable conditions for mosquito eggs and larvae.
- reanalysis_min_air_temp_k underwent a box cox transformation that improved its R value from 0.1882959 to 0.188747.

Some examples, from peer-reviewed journals of how the types of models we developed were used in similar situations are:

- In the peer-reviewed *Environmental Research* article "Time Series Regression Model for Infectious Disease and Weather" (Imai, Armstrong and Chalibi) the authors discuss the use of general linearized models and autoregressive integrated moving average (ARIMA) models to examine associations between environmental predictors and infectious disease.
- The journal of *Epidemiology and Infection* article "Social contacts of school children and the transmission of respiratory-spread pathogens" (Mikolajczyk, Akmatov and Rastin) article describes the use of negative binomial regression models to predict the spread of pathogens amongst school children.
- The article "Evaluating the links between climate, disease spread, and amphibian declines" (Rohr, Raffel and Romansic) discusses the use of general linearized models, to include Poisson and negative binomial, to associate human alteration of the climate with the decline of amphibian species.
- In the peer-reviewed *Acta Tropica* article (Pei-Chiu, How-Ran and Shih-Chun) "Weather as an effective predictor for occurrence of dengue fever in Taiwan" the authors evaluate the impact of weather variability and the occurrence of Dengue Fever in a major metropolitan Taiwanese city using autoregressive integrated moving average (ARIMA) models.
- The article "Use of Time Series Analysis in Infectious Disease Surveillance" (Allard) reviews the practical aspects of the use of ARIMA (autoregressive, integrated, moving average) modelling of time series as applied to the surveillance of reportable infectious diseases

Summary

During this analysis, we built multiple multivariate linear regression, Neural Network, and time series models utilizing Poisson, Negative Binomial regression and ARIMA with both untransformed and transformed (box cox transformation) variables to predict the total number of case of Dengue Fever for a given week in San Juan, Puerto Rico and Iquitos, Peru. Exploratory Data Analysis was conducted to understand the distributions and relationships between our regressor and response variables. We then implemented regressor variable Box Cox transformations to adhere to the assumptions of Linear Regression and improve the performance of our models. Finally, an iterative trail-and-error process using Mean Absolute Error to assess goodness-of-fit was utilize to compare the results of each model to our validation data and pick our selected model. Despite our initial hypothesis that more rain would contribute to increase cases of Dengue we learned that humidity and warmer temperatures were more significant to the spread of the virus. We also found that that more complicated models don't necessarily relate to better performance and sometimes simpler is better. If we had more time, we would like to further explore the performance of the Negative Binomial model over our other models, attempt additional variable transformations to further optimize our model, and implement dummy variables to incorporate the seasonality that we observed in the total_cases decomposed time series.

Works Cited

Allard, R. "Use of time-series analysis in infectious disease surveillance." *World Health Orgnization Bulletin* 76.4 (1998): 327-333.

Imai, Chisato, et al. "Time Series Regression Model for Infectious Disease and Weather." *Enviromental Research* 142 (2015): 319-327.

Mikolajczyk, et al. "Social Contacts of School Children and the Transmission of Respiratory Spread Pathogens." *Epidemology and Infection* 136.6 (2008): 813-822.

Pei-Chiu, Wu, et al. "Weather as an effective predictor for occurrence of dengue fever in Taiwan." *Acta Tropica* 103.1 (2007): 50-57.

Rohr, Jason, et al. "Evaluating the links between climate, disease spread, and amphibian declines." *Proceedings of the National Academy of the Sciences of the United States of America* 105.45 (2008): 17436-17441.