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

Daren Purnell

Predict 413 Section 55

August 26, 2017

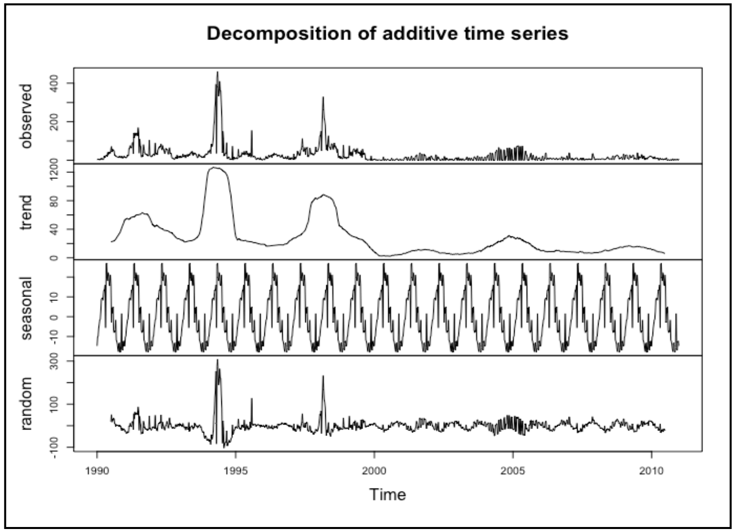
Executive Summary

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.

Exploratory data analysis

- Decomposition of the response variable, total_cases, time series.
  - Potential pandemics of Dengue occurring roughly around the early 1990s, 1994, and 1997.
  - There is a clear seasonal cycle of total cases that peaks during the summer months, with warmer weather, and declines as the weather cools.
  - 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.
- Computing the Pearson Correlation Coefficients of our data set allow us to better understand the linear relationship between the response and regressor variables .
  - Minimal impact from vegetation index and precipitation amount regressor variables.
  - Relatively significant linear relationships exist between temperature and humidity related variables.

Warmer, more humid, weather conditions contribute to the spread of Dengue Fever



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 (imputations and transformations)

Data preparation efforts consisted of variable transformation for ordinary least squares regression and imputation of missing records.

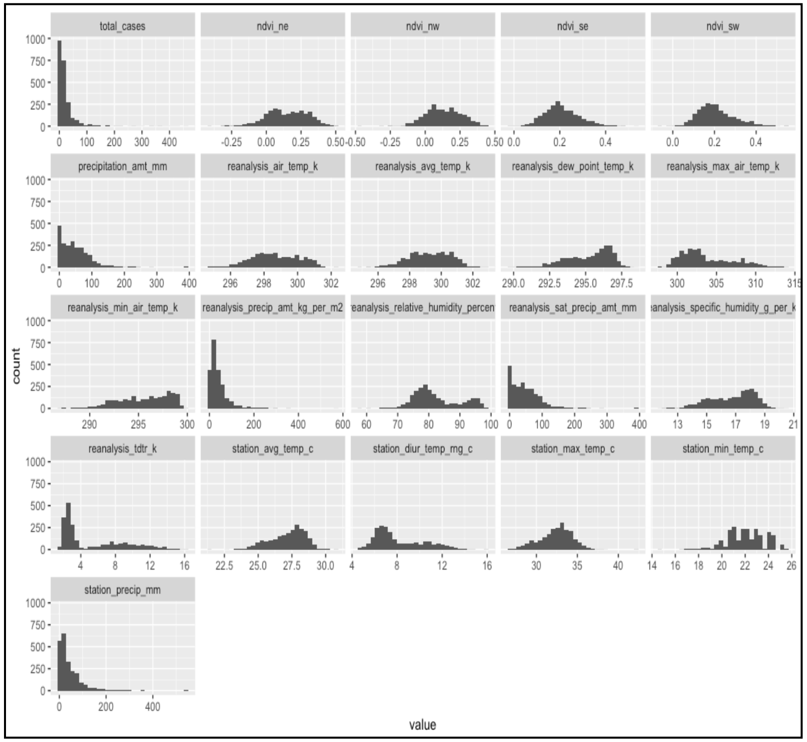
■While not necessary for generalized linear models, Box Cox variable transformation produced an improved predictive model over other methods of transformation.

■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.

■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.



Model development and selection

Development

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.

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

Selection

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
$\boldsymbol{\varepsilon}$	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.

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 implemented regressor variable Box Cox transformations to adhere to the assumptions of Linear Regression and improve the performance of our models.
- 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.
- Potential next steps:
  - Further explore the performance of the Negative Binomial model over our other models.
  - Attempt additional variable transformations to further optimize our model.
  - Implement dummy variables to incorporate the seasonality that we observed in the total_cases decomposed time series.