Crystal Mosley
Predict 413 – Time Series & Forecasting
Final – Summer 2018

Username: Mssbrightside25
Display Name: Crystal Mosley

DengAI: Predicting Disease Spread

Problem and Significance

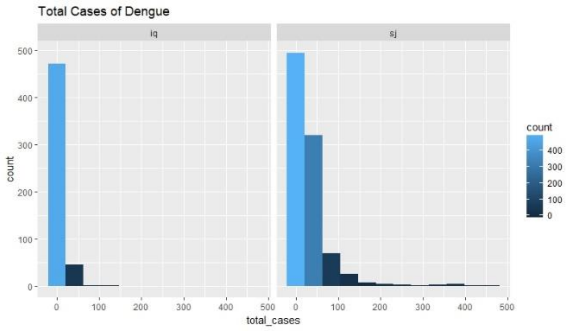
With more than one-third of the world's population living in areas at risk for infection, dengue virus is a leading cause of illness and death in the tropics and subtropics. As many as 400 million people are infected yearly. Dengue is caused by any one of four related viruses transmitted by mosquitoes. There are not yet any vaccines to prevent infection with dengue virus and the most effective protective measures are those that avoid mosquito bites. When infected, early recognition and prompt supportive treatment can substantially lower the risk of medical complications and death. Dengue has emerged as a worldwide problem only since the 1950s. Although dengue rarely occurs in the continental United States, it is endemic in Puerto Rico and in many popular tourist destinations in Latin America, Southeast Asia and the Pacific islands (CDC).

The goal of this competition from DrivenData's website is to predict local epidemics of Dengue Fever.

Data

To start this analysis of the data, I read in each of the provided datasets: *dengue_features_test, dengue_features_train, dengue_labels_train,* and *submission_format*. I reviewed each set and found there were only two cities listed: San Juan, Puerto Rico and Iquitos, Peru. I decided to separate the training data for both cities into their own sets of features and labels. I found there were nearly double the amount of entries for San Juan than for Iquitos – 936 entries for San Juan, and 520 entries for Iquitos (Figure 1).

Figure 1: Total_cases of Dengue for Iquitos (IQ) and San Juan (SJ)



Missing data was a big factor as there were 23 variables and a lot of these contained missing data with ndvi_ne missing the most – 194 out of 1436! There are a few options to handle this missing data: replacing the missing values with the most recently used, replacing with the mean or medium of each variable, or regression. I chose to replace the empty values with the previously used value. It did significantly change the vegetation index plot from looking normal except with gaps in the data (due to the NaNs), to looking odd with a low spike index (Figures 2 and 3).

Figure 2: Before Replacing Missing Values

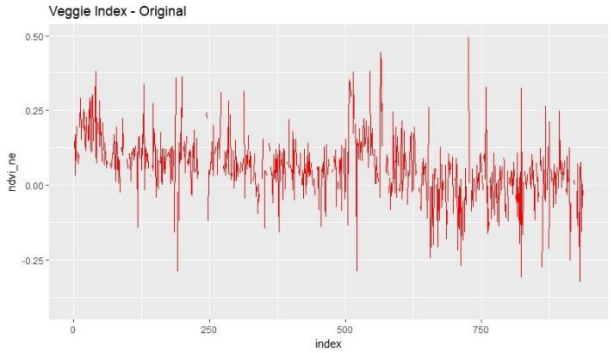
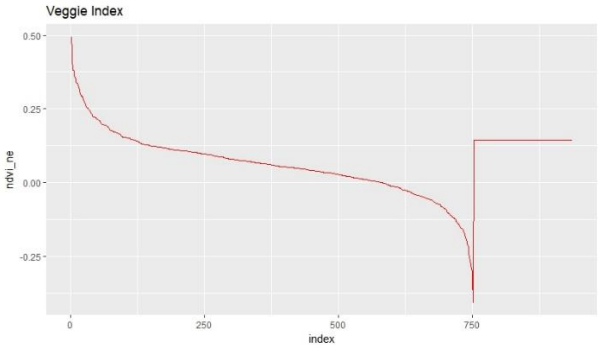
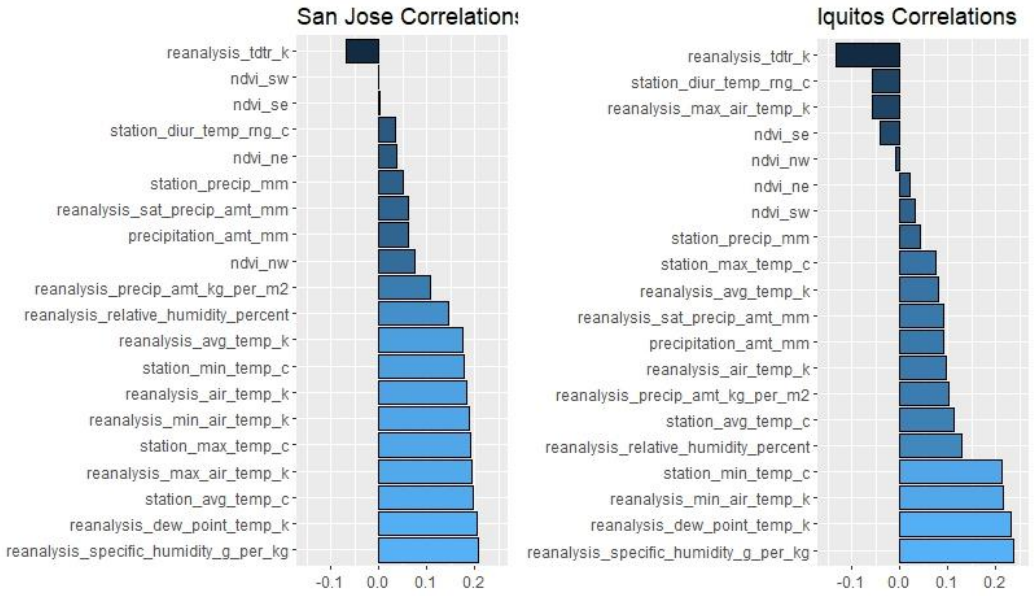


Figure 3: After Replacing Missing Values



The next task was to remove week_start_date as I didn't feel the need to use this as a feature and just focus on the climate variables and correlations. Following the analysis of the correlations between the predictor and response variables, the four strongest correlations to total_cases (response variable) were: reanalysis_specific_humidity_g_per_kg, reanalysis_dew_point_temp_k, station_avg_temp_c, and station_max_temp_c (Figure 4).

Figure 4: Correlation Histogram Plots of SJ and IQ



We see some volatility within the plots of the total cases of Dengue for both SJ and IQ, with IQ showing a little more (Figure 5). After gathering the information of both sets of training data and utilizing the ACF/PACF autocorrelations, there's 6 lags for SJ and 3 for IQ with both showing and looks to form an AR(1) model. I took the difference of each model and shows a better look at the trends and spikes in lags (Figure 6).

Figure 5: Total_cases volatility of SJ and IQ

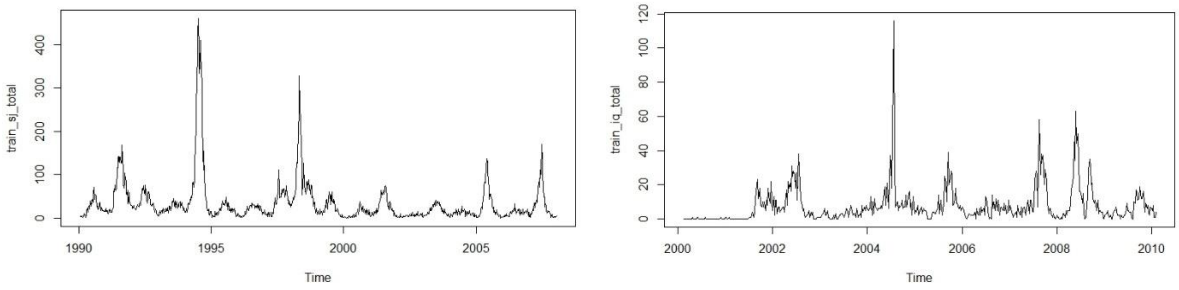
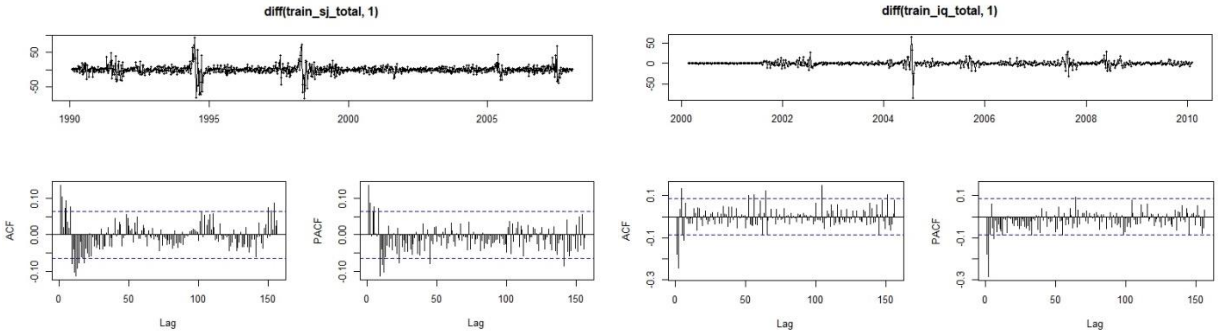


Figure 6: Lag differences of SJ and IQ



Literature Review

After researching other groups whom have conducted research in forecasting for disease epidemics, I found there are five types of models that are most commonly used: AR, MA, ARMA, ARIMA, and SARIMA. Back in 2013, a study was conducted on four different time series methods in forecasting Typhoid Fever incidence in China (NCBI). The used SARIMA and Neural Network models to work through their forecasting. SARIMA models combine seasonal differencing with an ARIMA model and are best when used with data that exhibit seasonal trend as they can effectively demo a linear model that can clutch the linear trend of the series. Another group utilized ANN (artificial neural networks) for their forecasting study on Brucellosis, which remains widespread globally in both domesticated animals and humans. This study used human brucellosis (HB) as a test case to identify important environmental determinants of the disease and predict its outbreaks. An artificial neural network (ANN) model was developed, using annual county-level numbers of HB cases and data on 37 environmental variables, potentially associated with HB in Inner Mongolia, China, due to how effective it is in forecasting disease outbreaks (MDPI). Another case where ANN's proved to be invaluable were detection of flu and

prediction through social media sites. The ANN analyzed flu-related and flu-symptoms-related keywords in Twitter. The extracted information was converted to flu-score using machine learning techniques (tbiomed). Random Forests showed as a capable model that outperforms ARIMA's in a lot of cases due to its enhanced predictive ability over existing time series models for the prediction of infectious disease outbreaks. Random Forests were used for prediction of avian flu H5N1 outbreaks back in 2014. The conclusion uncovered the time-series structure of outbreak severity for highly pathogenic avian flu in Egypt (NCBI). Random Forests are also good comparisons when trying to forecast disease epidemics with other models such as the new process model developed for classifying and forecasting epidemic curves: Dirichlet process - The DP model is a nonparametric Bayesian approach that enables the matching of current influenza activity to simulated and historical patterns, identifies epidemic curves different from those observed in the past and enables prediction of the expected epidemic peak time (europepmc).

Models / Formulation / Performance / Limitations

Model 1 – SARIMA with and without xreg (external regressors)

Seasonal ARIMA models are usually denoted ARIMA(p,d,q)(P,D,Q)m, where m refers to the number of periods in each season, and the uppercase P,D,Q refer to the autoregressive, differencing, and moving average terms for the seasonal part of the ARIMA model. The SARIMA model with xreg used only the correlated features listed in the *Data* section of this report. An ARIMA model was fit for SJ and parameters used order were (1,1,1) and seasonal (0,0,0). The residuals looked much better for the ACF/PACF plots as there were only a handful of lags and lower spikes. The "ljung" box test gave a p-value of 88% significance. On the other hand, the IQ model came out with a 95% p-value of significant when using parameters of order (0,1,2) and seasonal (0,0,1). Moving to SARIMA without xreg, all the features were used rather than just the related features. An ARIMA model was fit for SJ and parameters used order were (1,1,1) and seasonal (0,0,0). The residuals looked much better for the ACF/PACF plots as there were only a handful of lags and lower spikes. The "ljung" box test gave a p-value of 89% significance. On the other hand, the IQ model came out with a 95% p-value of significant when using parameters of order (0,1,2) and seasonal (0,0,1). The p-value was 1% better for this model (without xreg), but the RSME was a tiny bit higher than the model when using xreg.

Training Set	RMSE w/xreg	RMSE w/o xreg
San Juan	13.40983	13.42959
Iquitos	7.181565	7.191928

Score for both: 33.5024

Model 2 – Random Forest

A random forest is a multi-way classifier which consists of number of trees, with each tree grown using some form of randomization. The leaf nodes of each decision tree are labelled by estimates of the posterior distribution over the image classes. Each internal node contains a test that best splits the space of data to be classified. An image is classified by sending it down every

tree and aggregating the reached leaf distributions. For this model, all predictor variables were used, and 500 trees were created via regression random forest for both the SJ and IQ data. The mean of squared residuals for SJ was equal to 1953.313, and for IQ it was much lower, 110.3924.

Score: *Receiving this error when submitting – "IDs for submission are not correct."*

Model 3 – Neural Network

Neural networks (ANN) were designed to be modelled after the structure of the brain. ANN's consist of many hidden layers. Each hidden layer consists of multiple nodes. Each node is linked to other nodes using incoming and outgoing connections. Each of the connections can have a different, adjustable weight. Data is passed through these many hidden layers and the output is eventually interpreted as different results. The results used an average of 25 networks with 205 weights for the SJ set, and 145 for the IQ set. Just like the SARIMA models, the RSME was lower for the IQ dataset, but compared to the other predictive models, the RSME was lower overall when using the ANN.

Training Set	RMSE
San Juan	7.846459
Iquitos	3.680128

Score: 36.5938

Future Work

Massaging through the data a little more and diving deeper into EDA would help the model building process, but I felt that if I spent too much time going down that route, I would like time in trying to build out my models and go through the scoring process. There are other models that I would like to explore such as gradient boosting/xgboost, SVM's, RBF's for ANN – really, I would like to explore all the machine learning techniques on this data to see how the RMSE and scoring would affect the results.

Learning

I did not get a chance to model the midterm data due to starting so late, but this time I was able to create 3 models with 1 model having 2 variations in parameters. This is very new to me so getting experience with ARIMA/SARIMA, Random Forests, and Neural Networks in an actual predictive modeling process was a great learning opportunity.

Citations

Alessa, Ali, and Miad Faezipour1. “A Review of Influenza Detection and Prediction through Social Networking Sites.” *Theoretical Biology and Medical Modelling*, 1 Feb. 2018. https://tbiomed.biomedcentral.com/track/pdf/10.1186/s12976-017-0074-5 Accessed 27 August 2018.

Kane, Michael J, et al. *Advances in Pediatrics*., U.S. National Library of Medicine, 2014. www.ncbi.nlm.nih.gov/pmc/articles/PMC4152592/ Accessed 27 August 2018.

Nsoesie, EO, et al. “A Dirichlet Process Model for Classifying and Forecasting Epidemic Curves.” *Health Communication*, Oxford PharmaGenesis, Oxford. www.europepmc.org/articles/pmc3901791.com Accessed 27 August 2018.

Wang, Jiao, et al. “A Remote Sensing Data Based Artificial Neural Network Approach for Predicting Climate-Sensitive Infectious Disease Outbreaks: A Case Study of Human Brucellosis.” *MDPI*, 30 Sept. 2017. file:///C:/Users/crmo/Downloads/remotesensing-09-01018-v2.pdf Accessed 27 August 2018.

Zhang, Xingyu, et al. *Advances in Pediatrics*., U.S. National Library of Medicine, 2013, www.ncbi.nlm.nih.gov/pmc/articles/PMC3641111/ Accessed 27 August 2018.