Predicting Disease Spread

Scott Herman

August 28, 2017

Prepared for Predict-413: Time Series Analytics and Forecasting
Northwestern University Masters in Science, Predictive Analytics
Data-Driven File Submission: **rf_model_sth_nw.csv**

Problem

The purpose of this analysis is to develop a predictive model that will accurately predict the total number of Dengue cases per week and per year specific to two separate cities; San Juan and Iquitos. The dataset was collected from the DengAI: Predicting Disease Spread Driven Data Competition website, and contains a total of 1,456 observations across 25 variables. The objective is to utilize this historical data to develop three different predictive models which will deployed on a test data set and evaluated based upon each model's mean absolute error rate (MAE). Therefore, the overall goal is to minimize the MAE.

Significance

According to U.S. Centers for Disease Control,"Accurate dengue predictions would help public health workers. . . and people around the world take steps to reduce the impact of these epidemics." Therefore, accurately predicting when and where disease outbreaks could occur has the potential to save thousands of lives across the globe each year. Saving even one life is significant on its own, so the benefits of this type of analysis could be invaluable to the human population.

Data

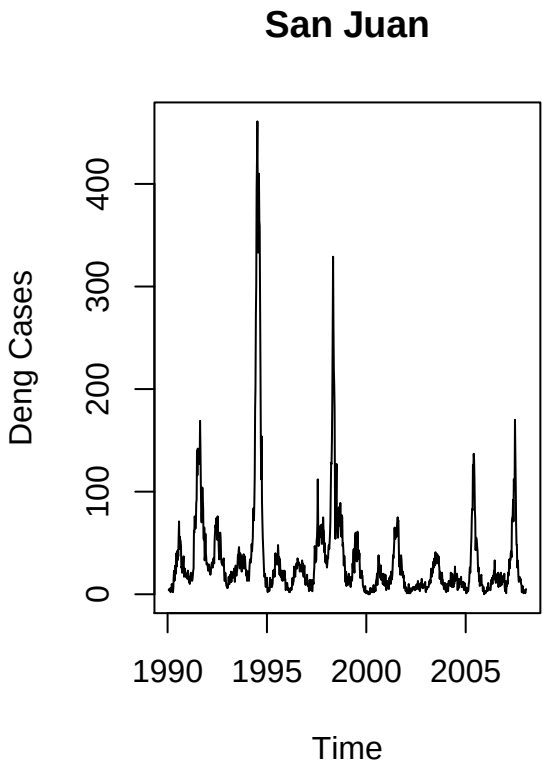
The table below gives the definitions of each variable within our initial data set. It appears that most of our variables are quantitiatve in nature, with the exception of city which is a factor variable. Our respose variable, target cases, appears to be an integer, which makes sense because each case would be recorded as a whole number. Additionally, there is a time series element that we'll want to pay attention to. Since we are trying to predict when a given outbreak occurs, we will want to ensure that our variables measuring units of time are coded correctly as Date/Time variables. This will allow us to create time series data frames and allow us to develop forecasts that pay special attention to the recorded dates in which Deng outbreaks are most likely to occur. Additionally, we did detect the presence of a total of 548 missing values that were contained by each of our predictor variables. These values were addressed by performing a linear interpolation to replace the missing values.

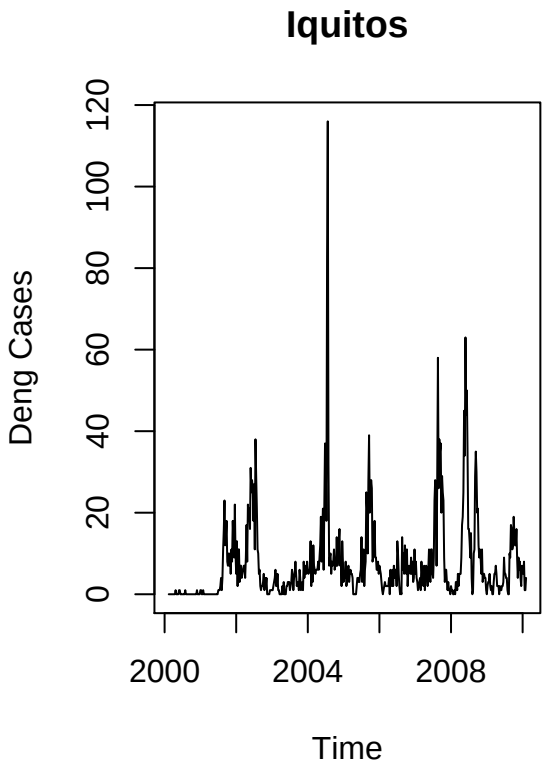
In terms of data prepartation, we felt it was necessary to split our training data in two different data frames based upon the city. In addition, we converted each of these data frames into time series object with the goal of increasing the accuracy of our models by allowing us to potentially better assess where and when a given outbreak is most likley to occur.

Table 1: Data Definitions

Variable	Description
city	City abbreviations: sj for San Juan and iq for Iquitos
year	Calendar Year ranging from 1990 to 2010
weekofyear	Recorded week of year
week_start_date	Date as Factor given in yyyy-mm-dd
ndvi_ne	Pixel northeast of city centroid
ndvi_nw	Pixel northwest of city centroid
ndvi_se	Pixel southeast of city centroid
ndvi_sw	Pixel southwest of city centroid
precipitation_amt_mm	Total precipitation
reanalysis_air_temp_k	Mean air temperature
reanalysis_avg_temp_k	Average air temperature
reanalysis_dew_point_temp_k	Mean dew temperature point
reanalysis_max_air_temp_k	Maximum air temperature
reanalysis_min_air_temp_k	Minumum air temperature
reanalysis_precip_amt_kg_per_m2	Total precipitation
reanalysis_relative_humidity_percent	Mean relative humidity
reanalysis_sat_precip_amt_mm	Total precipitation
reanalysis_specific_humidity_g_per_kg	Mean specific humidity
reanalysis_tdtr_k	Diurnal temperature range
station_avg_temp_c	Average temperature
station_diur_temp_rng_c	Diurnal temperature range
station_max_temp_c	Maximum temperature
station_min_temp_c	Minumum temperature
station_precip_mm	Total precipitation
total_cases	Recorded cases of Deng

After preparing our data, the figure below displays the plots of our two time series objects side-by-side. This reveals a few quick takeaways. First, we have two separate ranges of time that recorded the total cases specific to each city. The San Juan data was observed from 1990-2008, while the Iquitos data was recorded over a shorter period of time ranging from 2000-2010. Additionally, we can see that there are periods with huge spikes in the recorded number of cases without any clear overall trend identified in San Juan or Iquitos. Having said that, we also performed a decompositon of both time series objects, which suggest that the cases reported in Iquitos does show an element of seasonality. The Decomposition plots are given in the appendix for reference.





Model Development

In proceeding with our model development, we built three separate models attemting to predict the number of cases of Deng:

Neural Network

The first model develop was a Neural Network, which is an advanced modeling technique which develops forecasts based upon simple mathematical models that allow complex nonlinear relationships between the response variable and the set of predictors. This model was created by developing two separate Nueral Network models on each of the San Juan and Iquitos time series data frames. 25 networks were trained and the predictions were averaged for each model. This model was created in R using the format: fit_nnetar_iq <- nnetar(iq.ts,repeats=25, size=18, decay=0.1,linout=TRUE)

Arima Model

An Arima model was developed to due the fact that they are appropriate for time series forecasting and aim to describe autocorrelations found within the data. This model was

developed by creating two separate models on each city's time series data frame. Although we did not identify seasonality within the San Juan data, we did decide to include a seasonal component within each of these models. The results from each model were combined together to form the predictions for this model. These models were developed in R using the format: fit_iq <- Arima(iq.ts, order=c(0,1,2), seasonal=c(0,2,1))

Random Forest

The last model developed utilized a slightly different approach, and omitted the time series objects previously created. However, we did develop two Random Forest Models on two separate data sets specific to each city that included the entire set of predictors for both San Juan and Iquitos. The results from each model were combined to form our predictions. The R code used to develop these models used this type of format: iq_rf_model <- randomForest(total_cases ~. , data = iq)

Model Performance

After developing each model, we created our predictions on the test data set and submitted our final predictions to the Driven Data competition. The scores are summarized, below:

Table 2: Model Performance Summary

Model	Driven Data Score
NNetar	40.1923
Arima	33.9063
Random Forest	28.1418

As you can see above, the Random Forest model achieved the lowest MAE score of all three of our models, with a score of 28.1418. This was a bit surprising due to the fact that this problem was time series focused, yet this was the only model that did not utilize the time series objects to develop our predictions. Though, we expect that splitting up our data specific to each city did help improve these results. The Arima model results were second to the Random Forest model, while the Nueral Network resulted in the highest MAE and lowest level of accuracy.

Limitations and Future Work

A predictive model is only as good as its predictions, and in this case, there is still room for improvement. These results could be improved by:

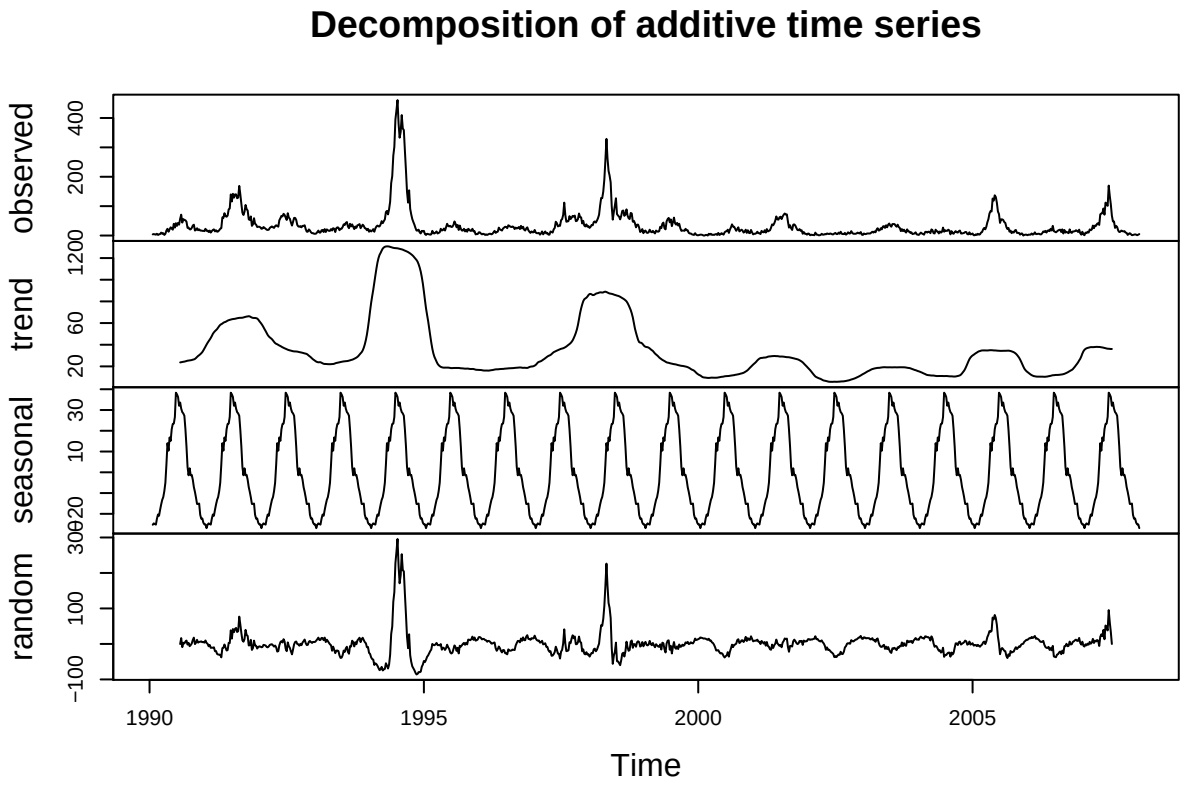
- Omitting variables possessing multicollinearity from our predictor set
- Utilizing transformations on our predictor variables
- Creating new variables with potential for increased correlation to our response
- Partitioning separate training and validation sets to better evaluate model results before deploying on test data
- Devloping additional types of time series forecasting models that utilize STL decompositions and/or Exponential Smoothing
- Better identifying cyclical and seasonal trends within time series objects

Learning

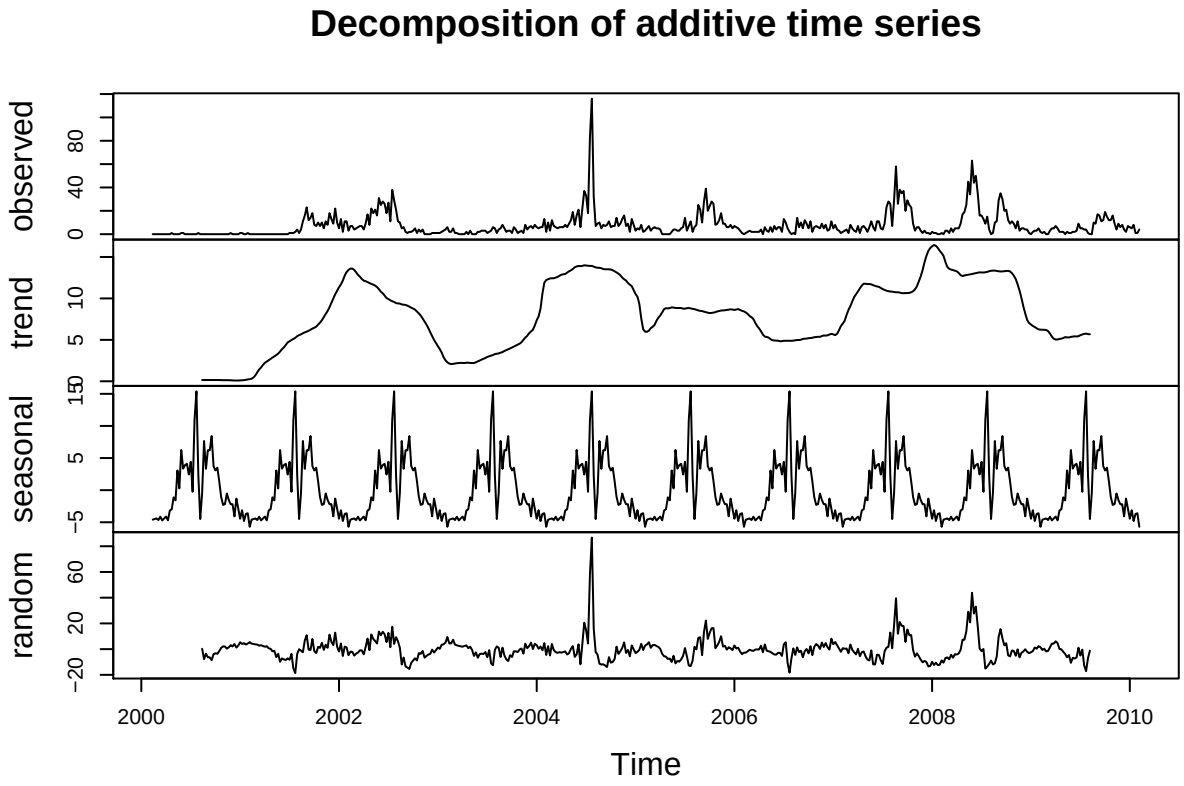
Attempting to solve this type of modeling problem was certainly a challenge. Despite developing time series models and forecasts all throughout this course, this assignment taught me that I still have a bit to learn. Although there are a wide variety of technqiues that can be utilized to build time series forecasts, being able to identify which type of procedure works best on the data at-hand will certainly be an on-going challenge. This is mainly due to the fact that each and every modeling challenge will be unique to the data you are given, and identifying the best way to maximize the model results requires time, practice, and patience.

Appendix

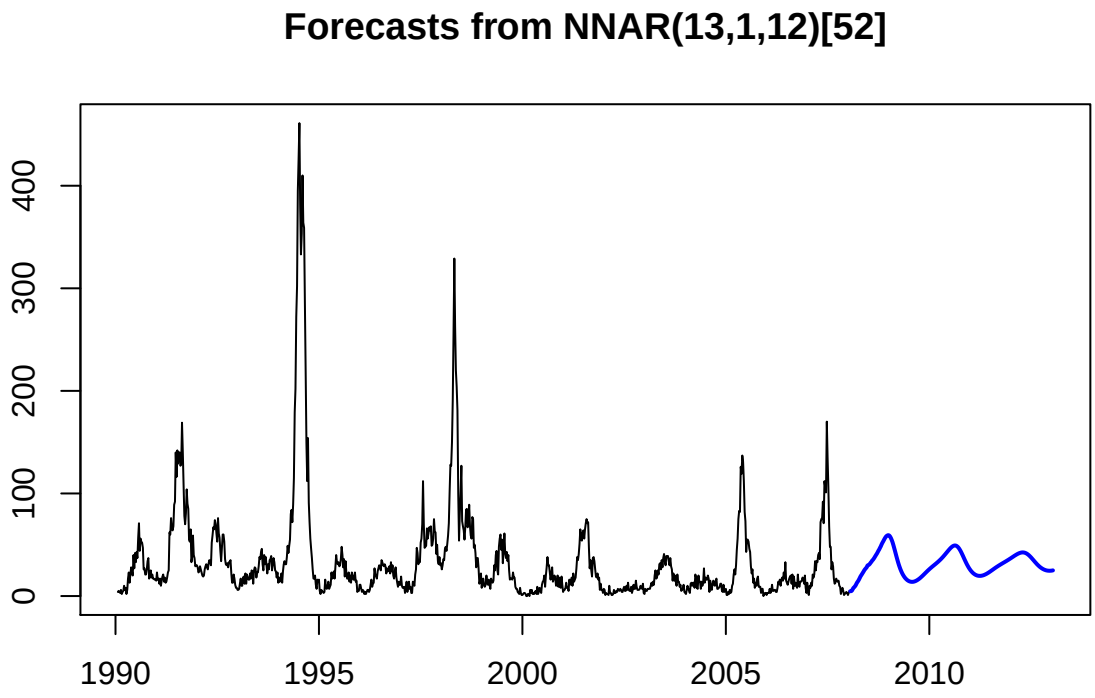
San Juan Time Series Decomposition



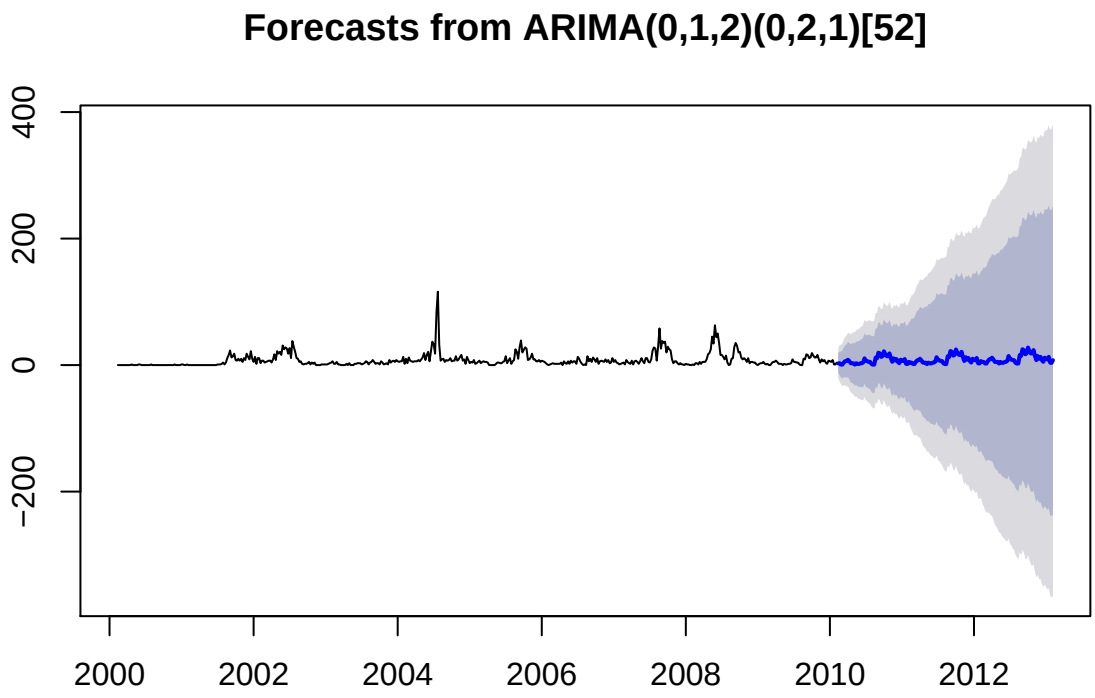
Iquitos Time Series Decomposition



Nural Network Forecast



Arima Forecast


Forecasts from ARIMA(0,1,2)(0,2,1)[52]
400
200
0
−200
2000
2002
2004
2006
2008
2010
2012