Final Project – Forecasting Dengue Spread

Time Series Analysis & Forecasting: MSCA 31006

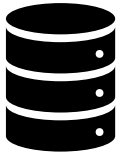
June, 2019

Team Members:

- Siddhartha Khetawat
- Srihari Seshadri

Introduction

Analyzing historical **Dengue cases** in San Juan, Puerto Rico to make weekly forecasts that will help the city be better prepared in the event of fatal epidemics

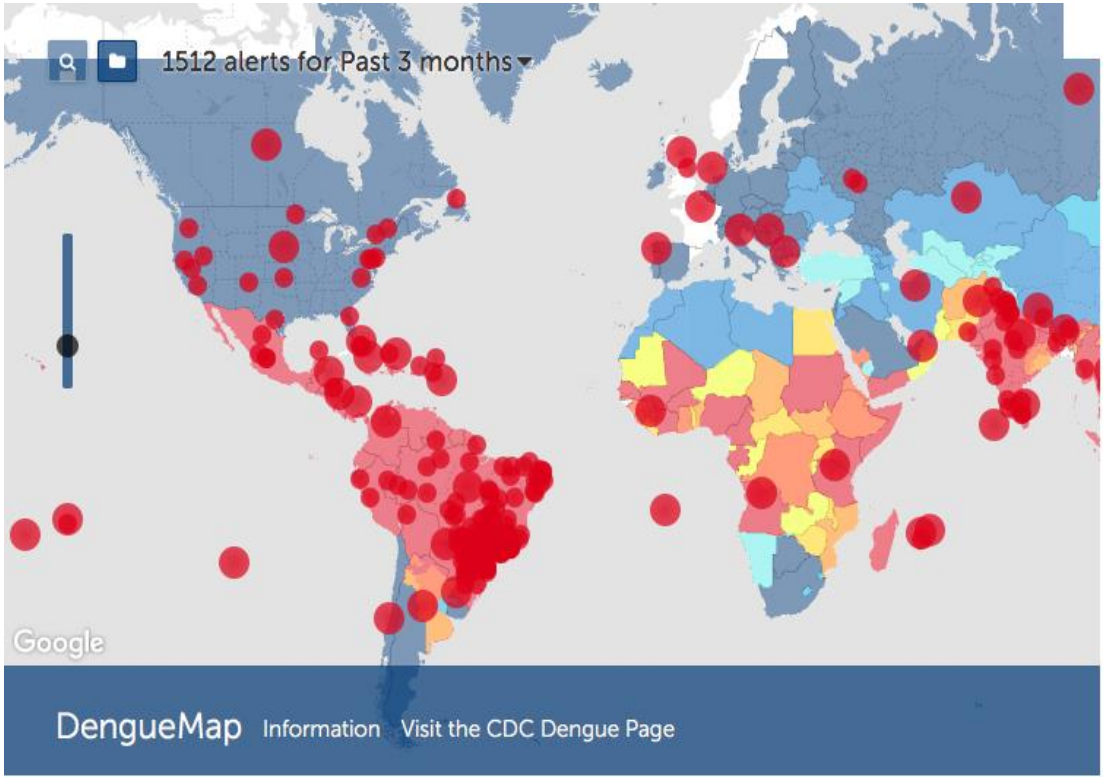


Using **environmental data** collected by Centers for Disease Control & Prevention and the National Oceanic & Atmospheric Administration in the U.S. Department of Commerce

Explored various **forecasting models** that account for components such as Trend and Seasonality and also external variables describing changes in temperature, precipitation, humidity, etc.

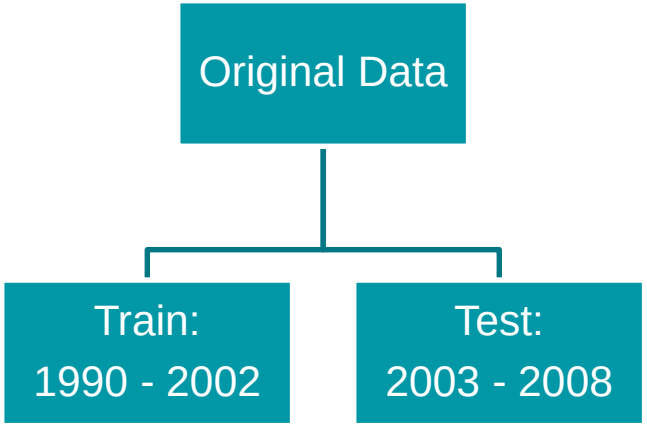
Problem Statement

- Dengue fever is a mosquito-borne disease that occurs in tropical and sub-tropical parts of the world that can cause bleeding, low blood pressure, and even death in severe cases
- Because it is carried by mosquitoes, the transmission dynamics of dengue are related to climate variables such as temperature and precipitation
- In recent years, dengue fever has been spreading. Historically, the disease has been most prevalent in Southeast Asia and the Pacific islands. These days many of the nearly half billion cases per year are occurring in Latin America

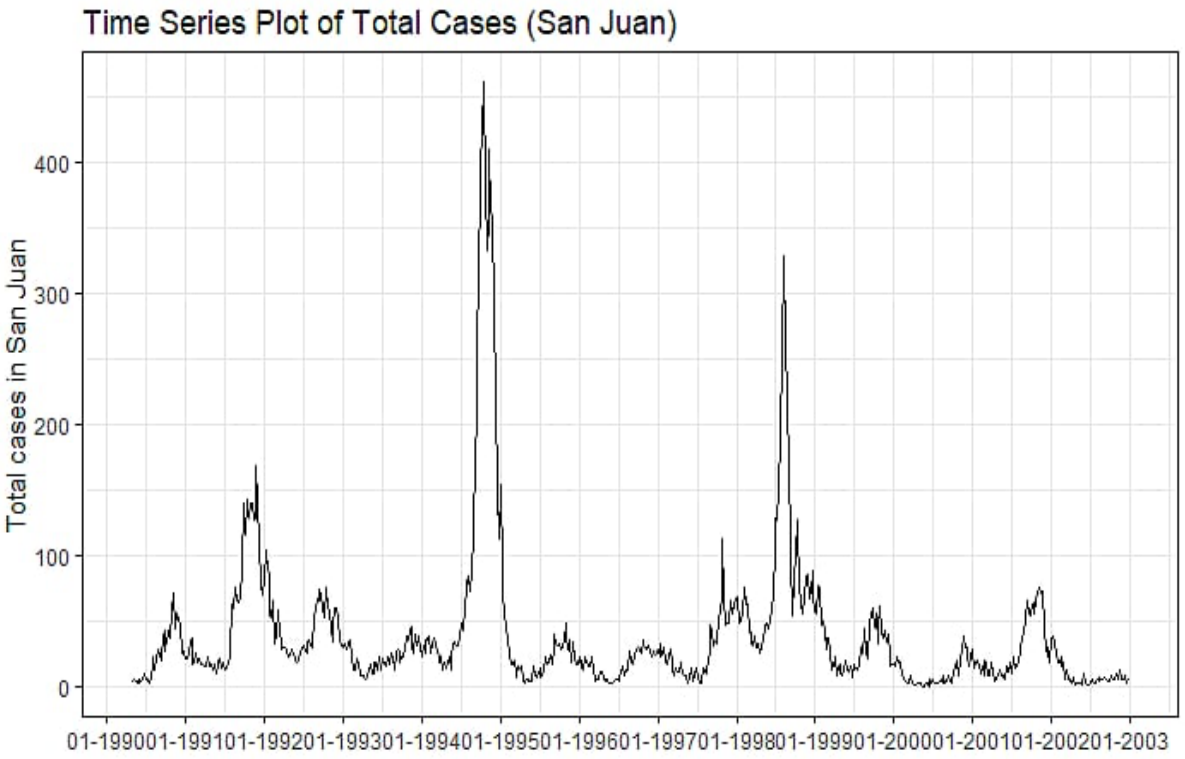


Data

```
Skim summary statistics
 n obs: 936
 n variables: 26

-- Variable type:character ---------------------------------------------------
 variable missing complete   n min max empty n_unique
    sep.x       0      936 936   1   1     0        1
    sep.y       0      936 936   1   1     0        1

-- Variable type:Date --------------------------------------------------------
       variable missing complete   n        min        max     median n_unique
 week_start_date       0      936 936 1990-04-30 2008-04-22 1999-04-26      936

-- Variable type:factor ------------------------------------------------------
 variable missing complete   n n_unique     top_counts ordered
     city       0      936 936        1 sj: 936, NA: 0   FALSE

-- Variable type:integer -----------------------------------------------------
    variable missing complete   n    mean    sd  p0    p25   p50   p75 p100     hist
 total_cases       6      930 936   34.34 51.5   0   9     19    37    461 ▇▁▁▁▁▁▁▁
  weekofyear       0      936 936   26.5  15.02  1   13.75 26.5  39.25  53 ▇▇▇▇▇▇▇▇
        year       0      936 936 1998.83  5.21 1990 1994  1999  2003  2008 ▇▆▆▆▆▆▅▇

-- Variable type:numeric -----------------------------------------------------
                              variable missing complete   n   mean    sd      p0     p25    p50    p75   p100     hist
                               ndvi_ne     191      745 936  0.058  0.11   -0.41   0.0045 0.058  0.11   0.49 ▁▁▂▇▂▁▁▁
                               ndvi_nw      49      887 936  0.067  0.092  -0.46   0.016  0.068  0.12   0.44 ▁▁▃▇▂▁▁▁
                               ndvi_se      19      917 936  0.18   0.057  -0.016  0.14   0.18   0.21   0.39 ▁▁▂▇▇▂▁▁
                               ndvi_sw      19      917 936  0.17   0.056  -0.063  0.13   0.17   0.2    0.38 ▁▁▂▇▇▂▁▁
                  precipitation_amt_mm       9      927 936 35.47  44.61   0       0      20.8   52.18 390.6 ▇▁▁▁▁▁▁▁
                 reanalysis_air_temp_k       6      930 936 299.16  1.24  295.94  298.2  299.25 300.13 302.2 ▁▁▂▃▇▇▃▂
                 reanalysis_avg_temp_k       6      930 936 299.28  1.22  296.11  298.3  299.38 300.23 302.16 ▁▁▂▃▇▇▃▂
            reanalysis_dew_point_temp_k      6      930 936 295.11  1.57  289.64  293.85 295.46 296.42 297.8 ▁▁▂▃▅▇▅▂
              reanalysis_max_air_temp_k      6      930 936 301.4   1.26  297.8   300.4  301.5  302.4  304.3 ▁▁▂▃▅▇▇▃
              reanalysis_min_air_temp_k      6      930 936 297.3   1.29  292.6   296.3  297.5  298.4  299.9 ▁▁▂▃▅▇▇▃
         reanalysis_precip_amt_kg_per_m2     6      930 936 30.47  35.63   0      10.83  21.3   37     570.5 ▇▁▁▁▁▁▁▁
   reanalysis_relative_humidity_percent      6      930 936 78.57   3.39  66.74   76.25  78.67  80.96  87.58 ▁▁▂▃▅▇▅▂
 reanalysis_specific_humidity_g_per_kg       6      930 936 16.55   1.56  11.72   15.24  16.85  17.86  19.44 ▁▁▂▃▅▇▅▂
                      reanalysis_tdtr_k      6      930 936  2.52   0.5    1.36    2.16   2.46   2.8    4.43 ▁▁▂▃▅▇▃▂
                    station_avg_temp_c       6      930 936 27.01   1.42  22.84   25.84  27.23  28.19  30.07 ▁▁▂▃▅▇▅▂
               station_diur_temp_rng_c       6      930 936  6.76   0.84   4.53    6.2    6.76   7.29   9.91 ▁▁▂▃▅▇▃▂
                    station_max_temp_c       6      930 936 31.61   1.72  26.7    30.6   31.7   32.8   35.6 ▁▁▂▃▅▇▅▂
                    station_min_temp_c       6      930 936 22.6    1.51  17.8    21.7   22.8   23.9   25.6 ▁▁▂▃▅▇▅▂
                     station_precip_mm       6      930 936 26.79  29.33   0       6.82  17.75  35.45 305.9 ▇▁▁▁▁▁▁▁
```



Data Imputation



Used **LOCF** (Last Observation Carried Forward) to impute these missing observations in the data

Example of missing data

year <int>	weekofyear <int>	total_cases <int>
1990	28	NA
1992	15	NA
1993	14	NA
1995	20	NA
1996	22	NA
1997	19	NA

Assumptions – Stationarity

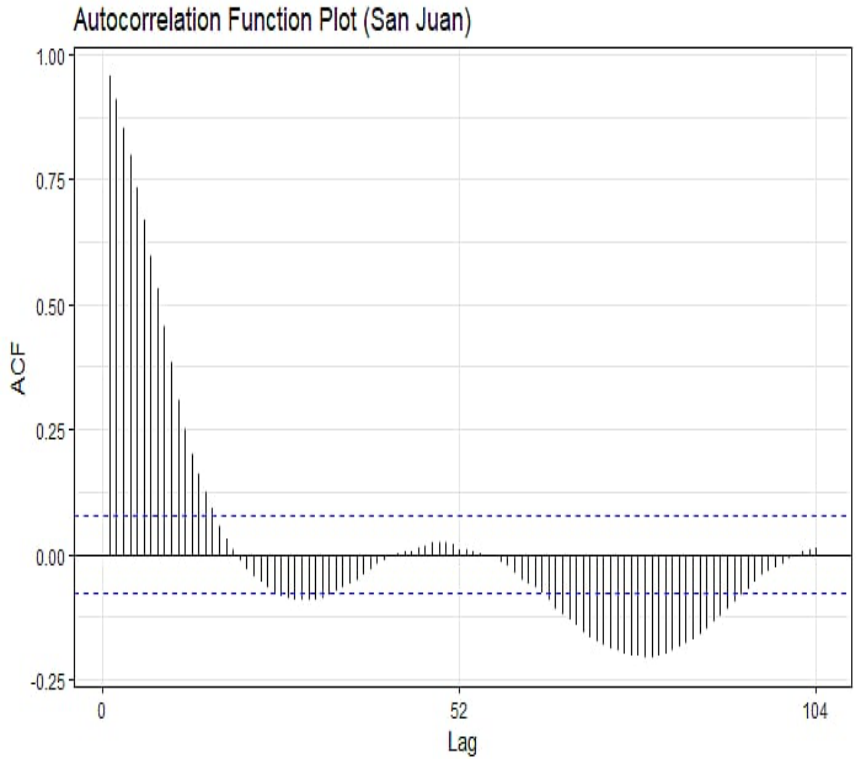
ADF Test:

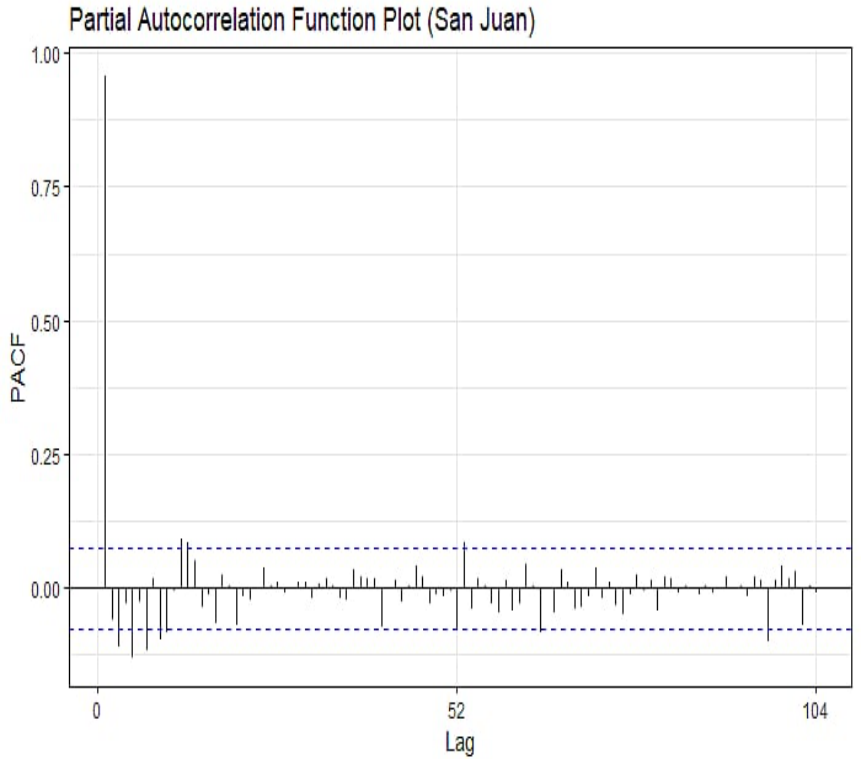
```
        Augmented Dickey-Fuller Test

data:  sjdata.train$total_cases
Dickey-Fuller = -5.8906, Lag order = 8, p-value = 0.01
alternative hypothesis: stationary
```

KPSS Test:

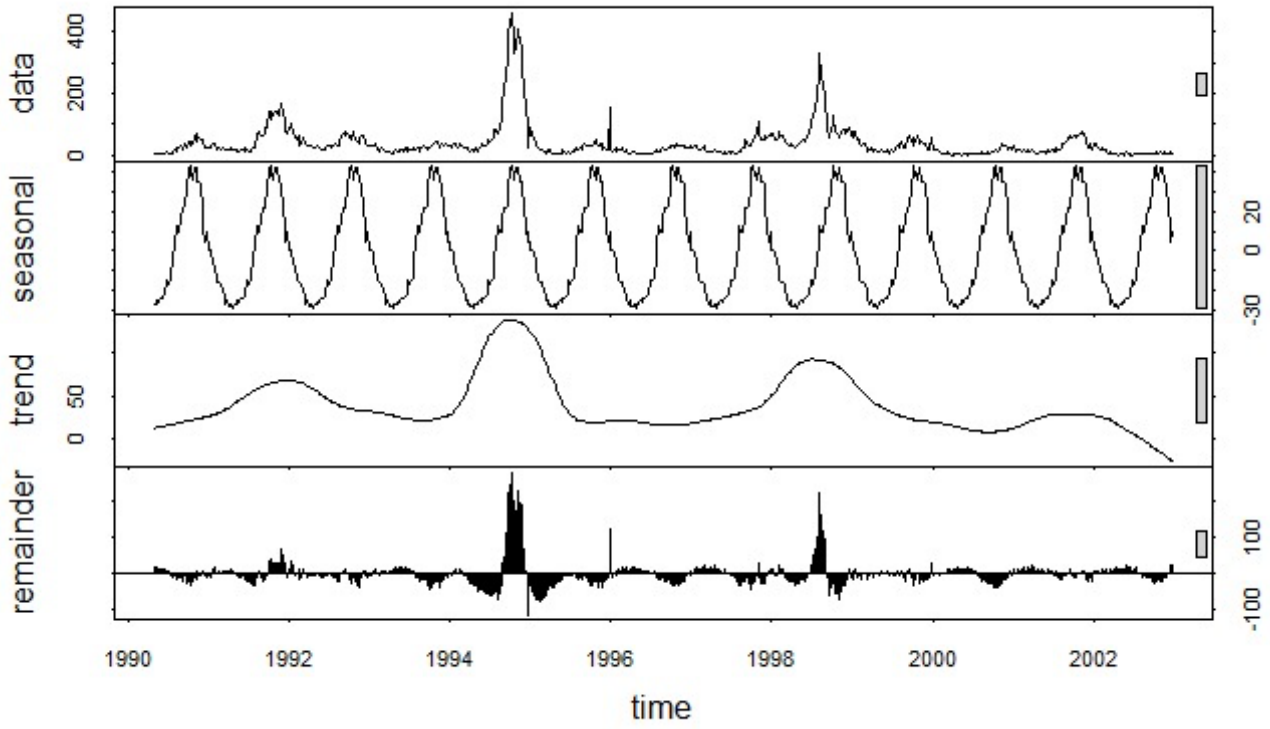
```
        KPSS Test for Level Stationarity

data:  sjdata.train$total_cases
KPSS Level = 0.3684, Truncation lag parameter = 6, p-value = 0.09078
```

Assumptions – ACF / PACF





Data Properties – Seasonality / Trend

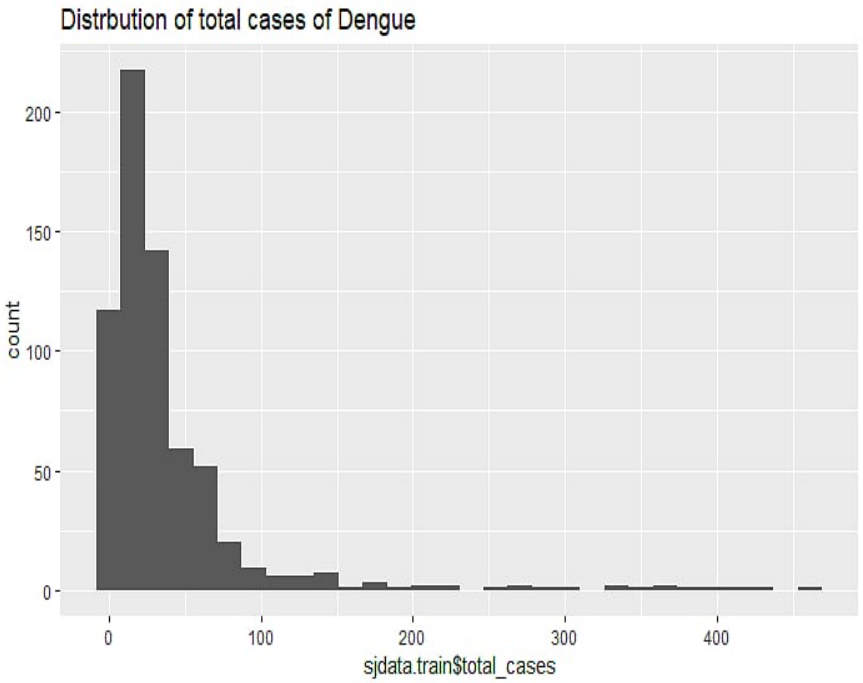


Observations:

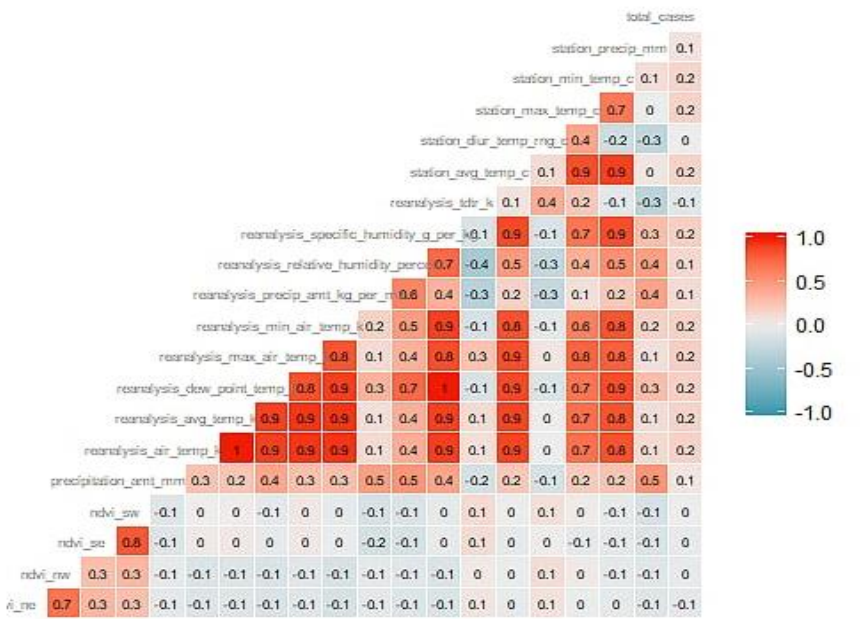
- We break down the “number of cases” time series into Seasonality, Trend and Remainder
- Looking at the Seasonality plot, we can clearly see a seasonal pattern with peaks observed every year around summer time
- We can observe some levels in the Trend plot, with both positive and negative trend occurring over the years

Exploratory Data Analysis

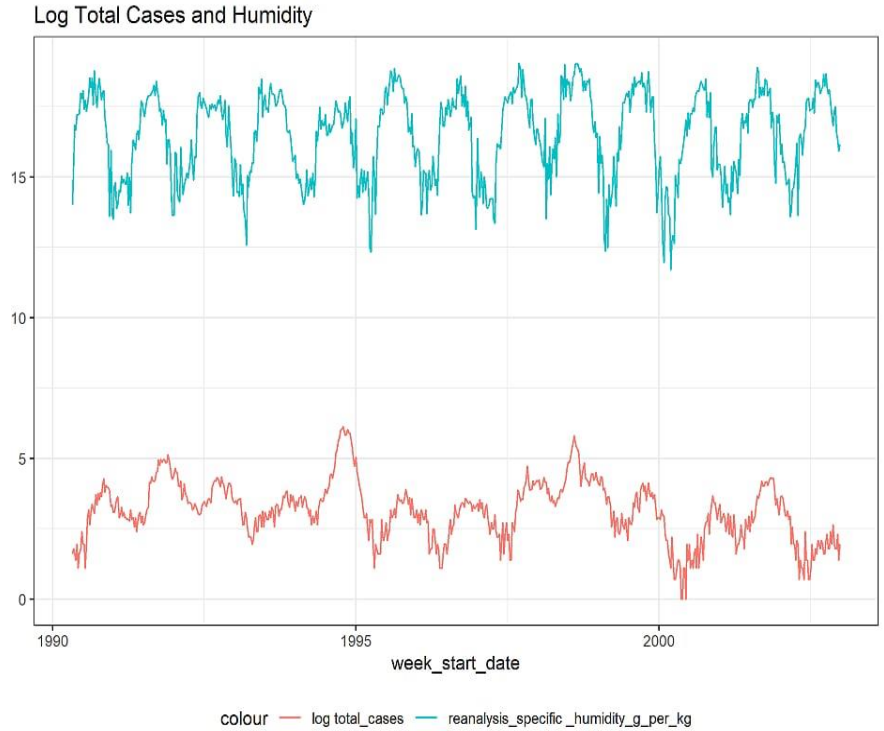
Histogram

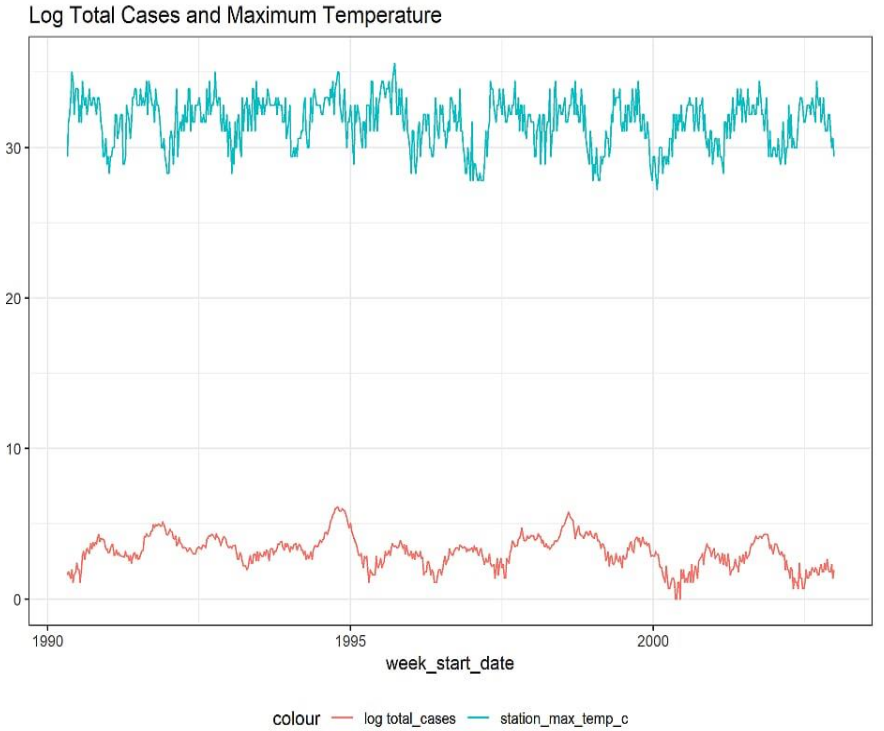


Correlations



Exploratory Data Analysis





Model 1 – ARIMA

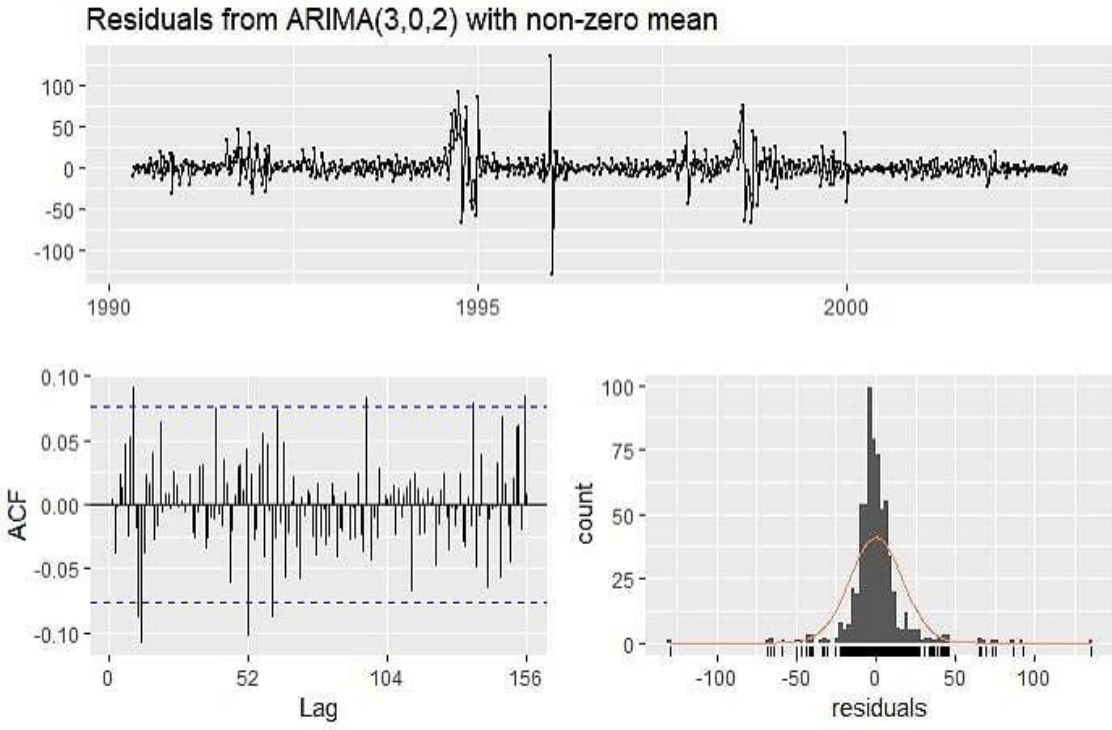
Model Summary:

```
Series: sjtrain.ts
ARIMA(3,0,2) with non-zero mean

Coefficients:
         ar1     ar2      ar3     ma1      ma2     mean
      0.9208  0.8686  -0.8161  0.0720  -0.7303  38.9242
s.e.  0.0601  0.0767   0.0543  0.0743   0.0707   8.2134

sigma^2 estimated as 276:  log likelihood=-2785.33
AIC=5584.67   AICc=5584.84   BIC=5616.1
```

ACF Test:

```
        Ljung-Box test

data:  Residuals from ARIMA(3,0,2) with non-zero mean
Q* = 96.623, df = 98, p-value = 0.5204

Model df: 6.   Total lags used: 104
```

Residual Analysis:



Model 2 – SARIMA

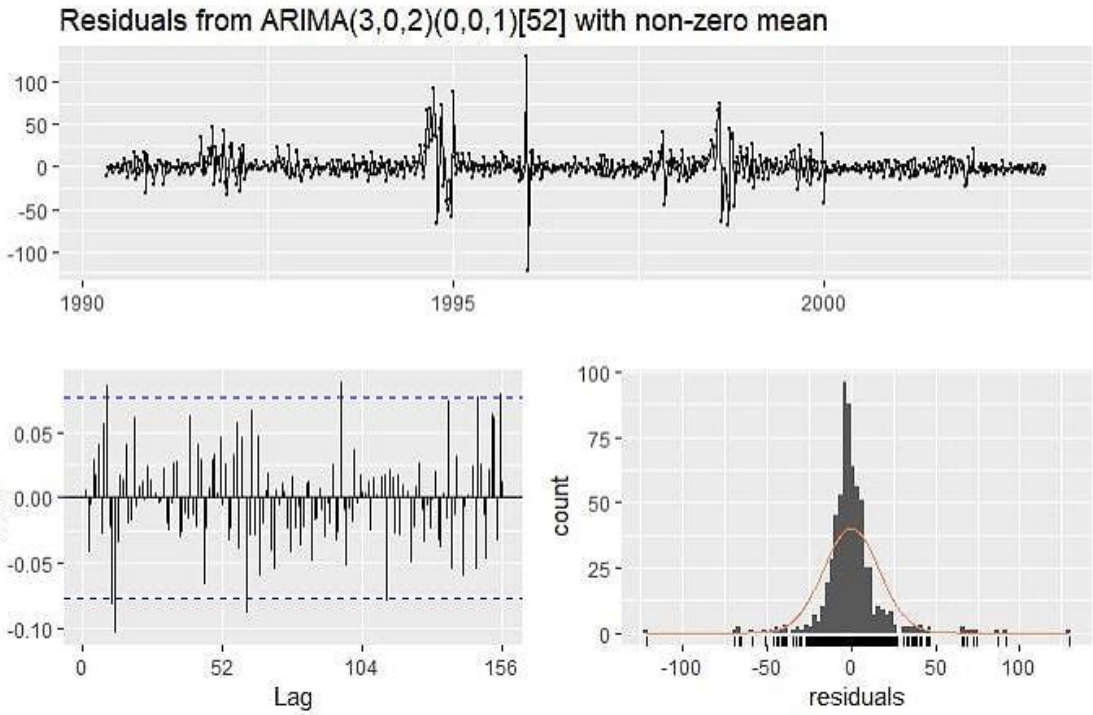
Model Summary:

```
Series: sjtrain.ts
ARIMA(3,0,2)(0,0,1)[52] with non-zero mean

Coefficients:
         ar1     ar2      ar3     ma1      ma2     sma1     mean
      0.9226  0.8788  -0.8282  0.0761  -0.7418  -0.1004  39.3919
s.e.  0.0570  0.0763   0.0522  0.0715   0.0686   0.0382   7.2312

sigma^2 estimated as 273.3:  log likelihood=-2781.92
AIC=5579.85   AICc=5580.07   BIC=5615.77
```

ACF Test:

```
        Ljung-Box test

data:  Residuals from ARIMA(3,0,2)(0,0,1)[52] with non-zero mean
Q* = 88.664, df = 97, p-value = 0.7151

Model df: 7.   Total lags used: 104
```

Residual Analysis:



Model 3 – SARIMAX

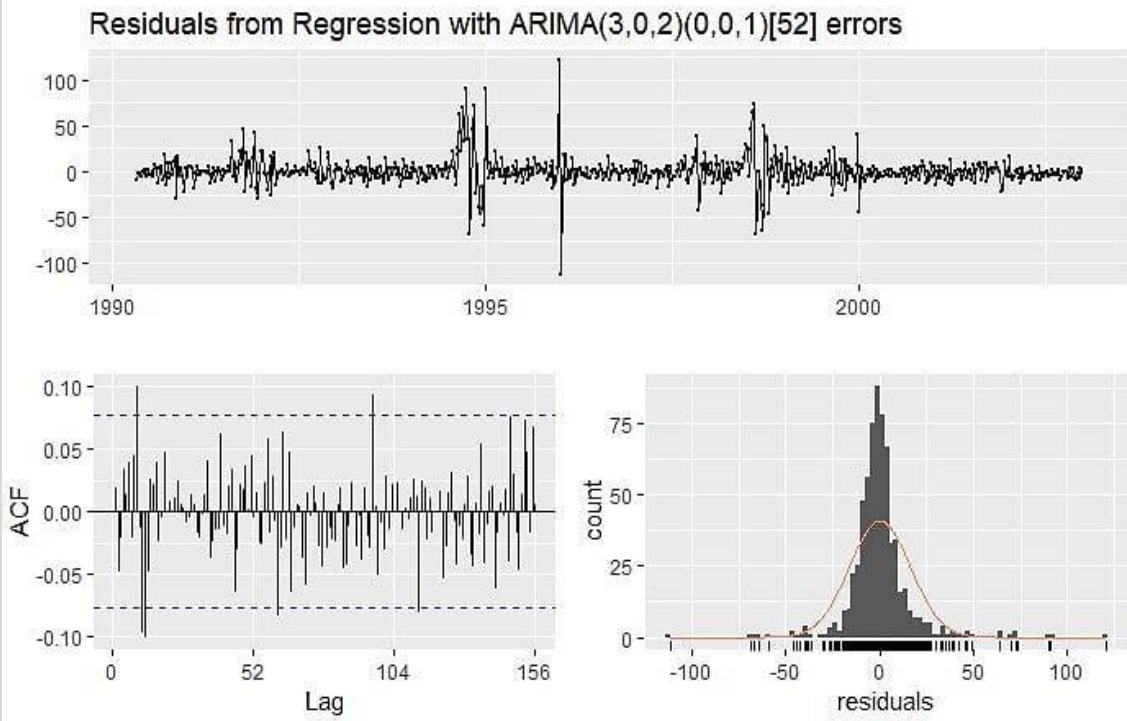
Model Summary:

```
Series: sjtrain.ts
Regression with ARIMA(3,0,2)(0,0,1)[52] errors

Coefficients:
         ar1     ar2      ar3     ma1      ma2     sma1  precipitation[, 1]
relative.humidity[, 1]  dew.temp[, 1]  avg.temp[, 1]  diur.temp[, 1]
      0.9109  0.8935  -0.8323  0.0909  -0.7407  -0.1005             -0.0025
0.0223         -0.1753          4.3603           0.5835
s.e.  0.0529  0.0599   0.0487  0.0676   0.0655   0.0388              0.0137
0.2363          0.1100          1.6925           0.9835
      max.temp[, 1]  min.temp[, 1]
             0.0055        -1.4431
s.e.         0.7211         0.9264

sigma^2 estimated as 268.6:  log likelihood=-2773.17
AIC=5574.35   AICc=5575   BIC=5637.22
```

ACF Test:

```
        Ljung-Box test

data:  Residuals from Regression with ARIMA(3,0,2)(0,0,1)[52] errors
Q* = 90.464, df = 91, p-value = 0.4961

Model df: 13.   Total lags used: 104
```

Residual Analysis:



Model 4 – Neural Net

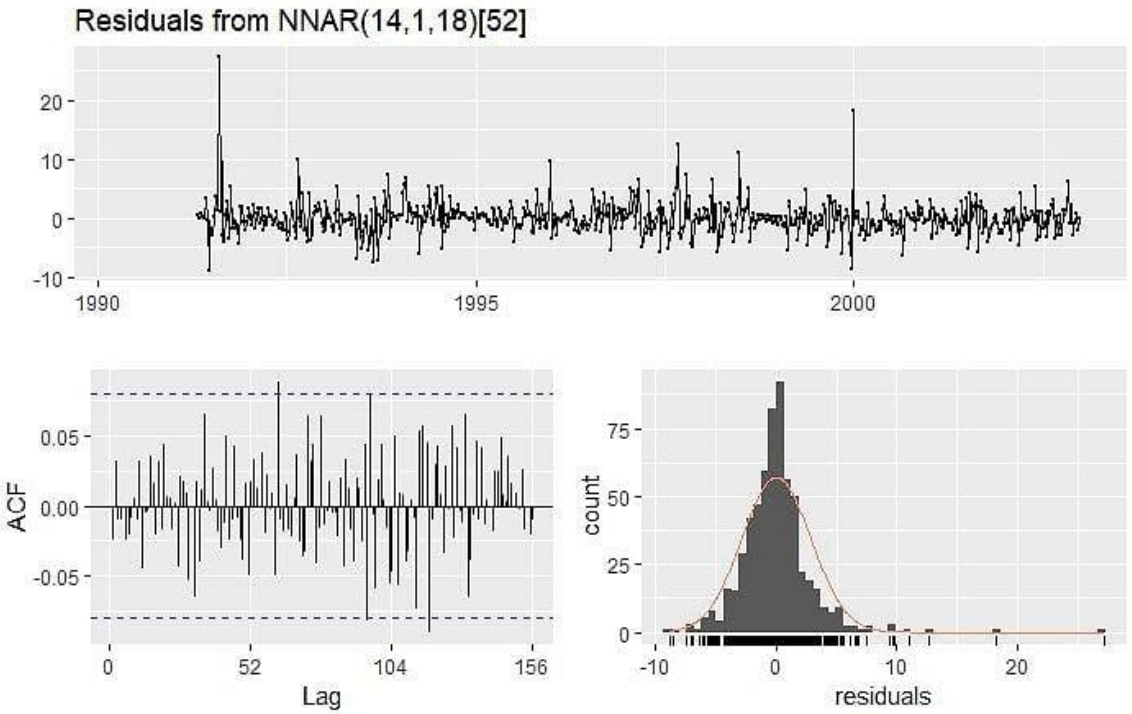
Model Summary:

```
Series: sjtrain.ts
Model:  NNAR(14,1,18)[52]
Call:   nnetar(y = sjtrain.ts, xreg = sjtrain.ts1[, 5:23])

Average of 20 networks, each of which is
a 34-18-1 network with 649 weights
options were - linear output units

sigma^2 estimated as 8.186
```

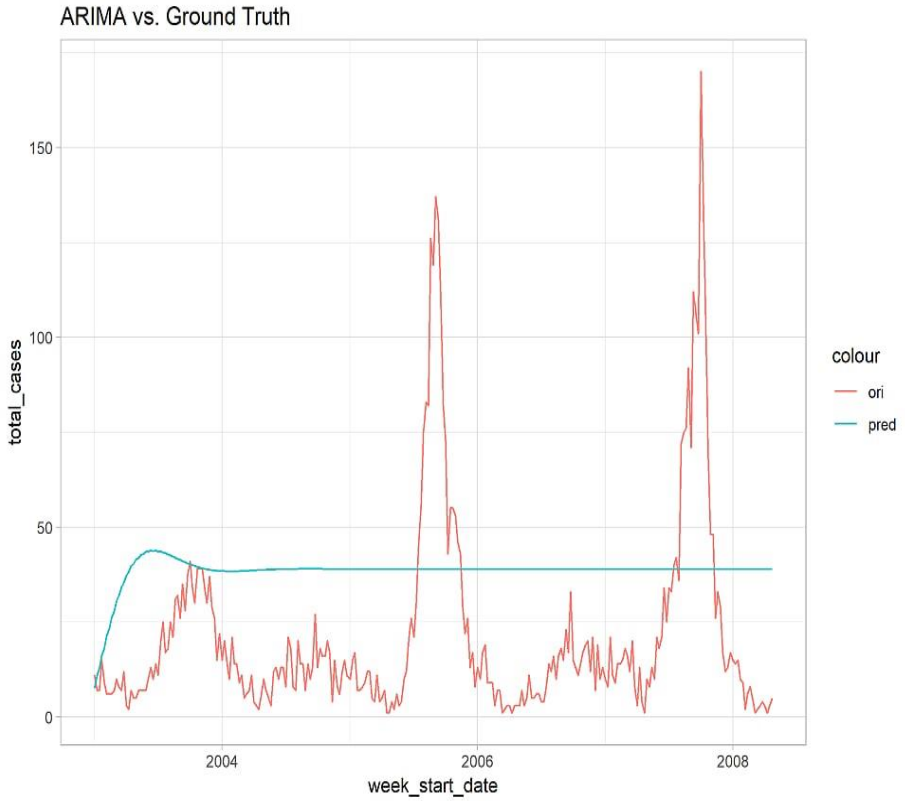
ACF Test:

```
        Box-Ljung test

data:  sj.fit4$residuals
X-squared = 0.35174, df = 1, p-value = 0.5531
```

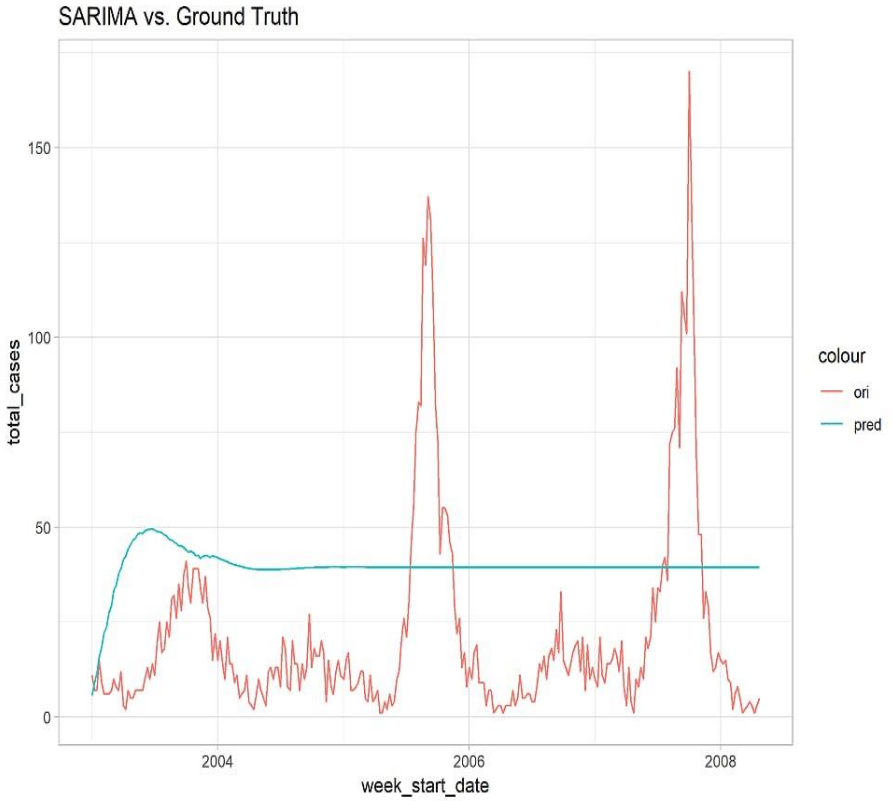
Residual Analysis:



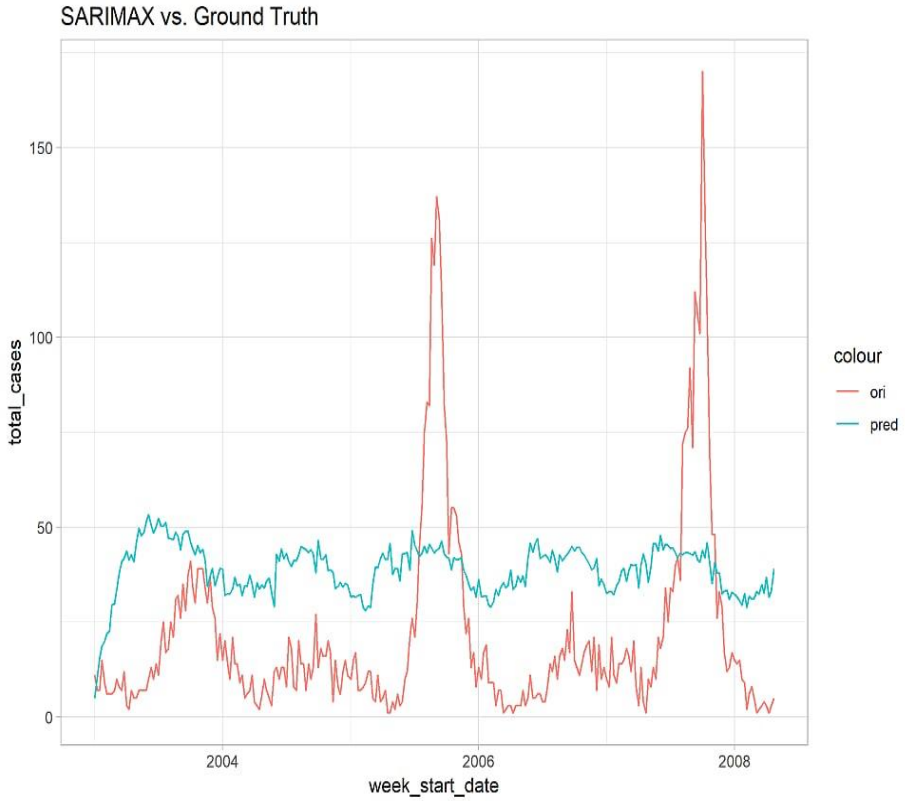
ARIMA Forecasts



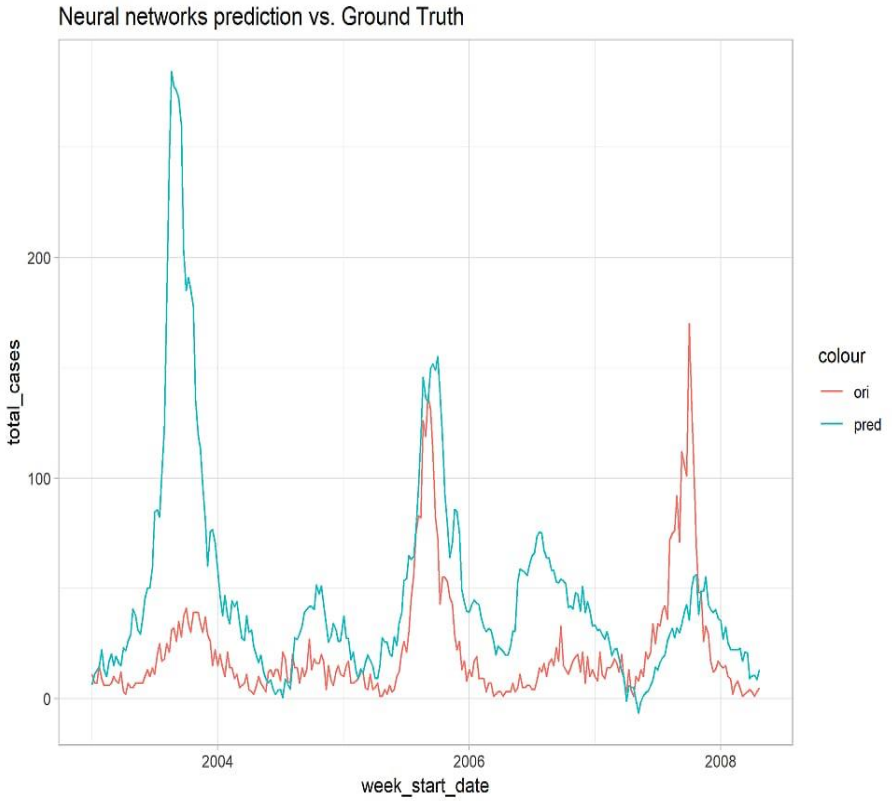
SARIMA Forecasts



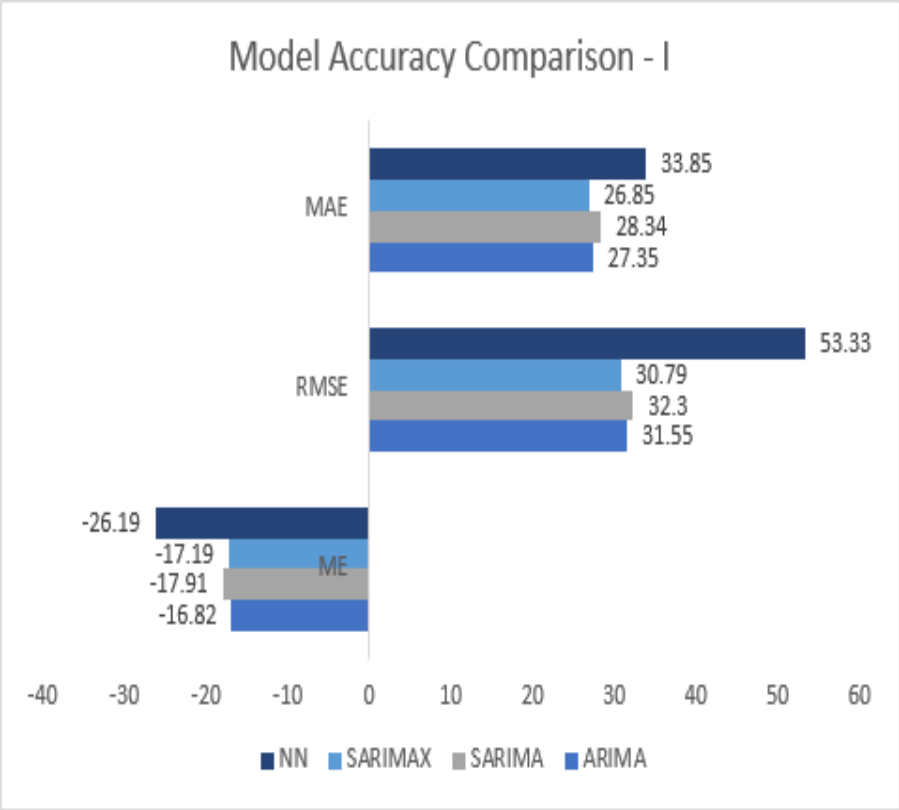
SARIMAX Forecasts

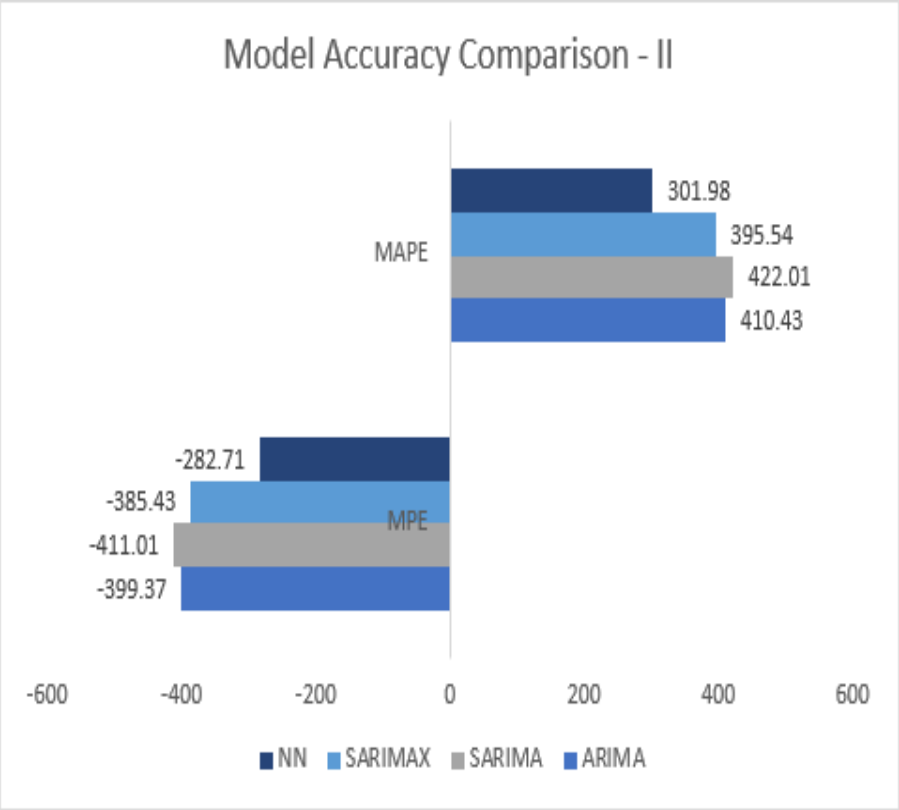


Neural Net Forecasts



Model Selection





Conclusion and Next Steps

Conclusion:

- Dengue fever cases have a seasonal pattern that can be modeled using various forecasting techniques
- We forecasted the test data using 4 models with increasing complexity using external variables
- SARIMAX and Neural Nets performed better on certain evaluation metrics because these models were able to accommodate environmental factors such as temperature, humidity and precipitation
- SARIMAX was able to capture the seasonality well, without being impacted by the anomalous data points
- The forecasts from Neural Nets was also pretty good but it got impacted by the anomalies in the data

Next Steps:

- Try some data transformations such as Box-Cox, Fourier transforms, etc. to better capture effects of predictors
- Develop an anomaly detection model to identify the outliers in the data which will help us improve the model
- Explore some other forecasting techniques such as Regression with ARMA errors and TBATS in case there is multiple seasonality in the data
- Explore deep learning models such as MPL, RNN and LSTM to get better forecasts