TimeSeriesProject

Srihari Seshadri

June 5, 2019

```
# Set working library
projects <- "D:/Dev/Sources/Projects/GitProjects/Predicting-Disease-Spread/"

# Shortcuts to folders of interest
CleanData <- paste0(projects,"/CleanData")
Dictionaries <- paste0(projects,"/Dictionaries")
RawData <- paste0(projects,"/RawData")
RCode <- paste0(projects,"/RCode")
RData <- paste0(projects,"/RData")
Output <- paste0(projects,"/Output")

# Load libraries and install them, if necessary
tmp.library.list <- c("haven", "zoo", "fUnitRoots", "tseries", "urca", "lmtest", "forecast", "data.table", "readxl","reshap
e", "quantmod", "ggplot2", "reshape2", "plyr","scales", "hts", "fpp2", "lubridate","stargazer","GGally")
for (i in 1:length(tmp.library.list)) {
  if (!tmp.library.list[i] %in% rownames(installed.packages())) {
    install.packages(tmp.library.list[i])
  }
  library(tmp.library.list[i], character.only = TRUE)
}
rm(tmp.library.list)
```

```
trainraw <- data.table(read.csv(paste0(RawData,"/dengue_features_train.csv"), sep=',', stringsAsFactors = F))
trainlraw <- data.table(read.csv(paste0(RawData,"/dengue_labels_train.csv"), sep=',', stringsAsFactors = F))
testraw <- data.table(read.csv(paste0(RawData,"/dengue_features_test.csv"), sep=',', stringsAsFactors = F))
submission <- data.table(read.csv(paste0(RawData,"/submission_format.csv"), sep=',', stringsAsFactors = F))

head(trainraw)
```

```
##    city year weekofyear week_start_date   ndvi_ne   ndvi_nw   ndvi_se
## 1:   sj 1990         18      1990-04-30 0.1226000 0.1037250 0.1984833
## 2:   sj 1990         19      1990-05-07 0.1699000 0.1421750 0.1623571
## 3:   sj 1990         20      1990-05-14 0.0322500 0.1729667 0.1572000
## 4:   sj 1990         21      1990-05-21 0.1286333 0.2450667 0.2275571
## 5:   sj 1990         22      1990-05-28 0.1962000 0.2622000 0.2512000
## 6:   sj 1990         23      1990-06-04        NA 0.1748500 0.2543143
##      ndvi_sw precipitation_amt_mm reanalysis_air_temp_k
## 1: 0.1776167                12.42              297.5729
## 2: 0.1554857                22.82              298.2114
## 3: 0.1708429                34.54              298.7814
## 4: 0.2358857                15.36              298.9871
## 5: 0.2473400                 7.52              299.5186
## 6: 0.1817429                 9.58              299.6300
##    reanalysis_avg_temp_k reanalysis_dew_point_temp_k
## 1:              297.7429                    292.4143
## 2:              298.4429                    293.9514
## 3:              298.8786                    295.4343
## 4:              299.2286                    295.3100
## 5:              299.6643                    295.8214
## 6:              299.7643                    295.8514
##    reanalysis_max_air_temp_k reanalysis_min_air_temp_k
## 1:                     299.8                     295.9
## 2:                     300.9                     296.4
## 3:                     300.5                     297.3
## 4:                     301.4                     297.0
## 5:                     301.9                     297.5
## 6:                     302.4                     298.1
##    reanalysis_precip_amt_kg_per_m2 reanalysis_relative_humidity_percent
## 1:                           32.00                              73.36571
## 2:                           17.94                              77.36857
## 3:                           26.10                              82.05286
## 4:                           13.90                              80.33714
## 5:                           12.20                              80.46000
## 6:                           26.49                              79.89143
##    reanalysis_sat_precip_amt_mm reanalysis_specific_humidity_g_per_kg
## 1:                        12.42                               14.01286
## 2:                        22.82                               15.37286
## 3:                        34.54                               16.84857
```

```
## 4:                       15.36                             16.67286
## 5:                        7.52                             17.21000
## 6:                        9.58                             17.21286
##    reanalysis_tdtr_k station_avg_temp_c station_diur_temp_rng_c
## 1:          2.628571           25.44286                6.900000
## 2:          2.371429           26.71429                6.371429
## 3:          2.300000           26.71429                6.485714
## 4:          2.428571           27.47143                6.771429
## 5:          3.014286           28.94286                9.371429
## 6:          2.100000           28.11429                6.942857
##    station_max_temp_c station_min_temp_c station_precip_mm
## 1:               29.4               20.0              16.0
## 2:               31.7               22.2               8.6
## 3:               32.2               22.8              41.4
## 4:               33.3               23.3               4.0
## 5:               35.0               23.9               5.8
## 6:               34.4               23.9              39.1
```

```
head(trainlraw)
```

```
##    city year weekofyear total_cases
## 1:   sj 1990         18           4
## 2:   sj 1990         19           5
## 3:   sj 1990         20           4
## 4:   sj 1990         21           3
## 5:   sj 1990         22           6
## 6:   sj 1990         23           2
```

```
dim(trainraw)
```

```
## [1] 1456   24
```

```
dim(trainlraw)
```

```
## [1] 1456    4
```

Merge the features data with the labels data

```
trainprep <- merge(trainraw, trainlraw, by=c("city", "year", "weekofyear"), all.x = T)
nrow(trainprep) == nrow(trainraw) # if true, we did not accidentally merge many-to-one
```

```
## [1] TRUE
```

```
dim(trainprep)
```

```
## [1] 1456   25
```

Omit reanalysis_sat_precip_amt_mm variable because of large amount of missing data

```
trainprep[, reanalysis_sat_precip_amt_mm := NULL]
```

Sort the dataset so it's in a clear order

```
setkeyv(trainprep, c("city", "year", "weekofyear"))

str(trainprep)
```

```
## Classes 'data.table' and 'data.frame':   1456 obs. of  24 variables:
##  $ city                                 : chr  "iq" "iq" "iq" "iq" ...
##  $ year                                 : int  2000 2000 2000 2000 2000 2000 2000 2000 2000 2000 ...
##  $ weekofyear                           : int  26 27 28 29 30 31 32 33 34 35 ...
##  $ week_start_date                      : chr  "2000-07-01" "2000-07-08" "2000-07-15" "2000-07-22" ...
##  $ ndvi_ne                              : num  0.193 0.217 0.177 0.228 0.329 ...
##  $ ndvi_nw                              : num  0.132 0.276 0.173 0.145 0.322 ...
##  $ ndvi_se                              : num  0.341 0.289 0.204 0.254 0.254 ...
##  $ ndvi_sw                              : num  0.247 0.242 0.128 0.2 0.361 ...
##  $ precipitation_amt_mm                 : num  25.4 60.6 55.5 5.6 62.8 ...
##  $ reanalysis_air_temp_k                : num  297 297 296 295 296 ...
##  $ reanalysis_avg_temp_k                : num  298 298 297 296 298 ...
##  $ reanalysis_dew_point_temp_k          : num  295 295 296 293 294 ...
##  $ reanalysis_max_air_temp_k            : num  307 307 304 304 307 ...
##  $ reanalysis_min_air_temp_k            : num  293 291 293 289 292 ...
##  $ reanalysis_precip_amt_kg_per_m2      : num  43.2 46 64.8 24 31.8 ...
##  $ reanalysis_relative_humidity_percent : num  92.4 93.6 95.8 87.2 88.2 ...
##  $ reanalysis_specific_humidity_g_per_kg: num  16.7 16.9 17.1 14.4 15.4 ...
##  $ reanalysis_tdtr_k                    : num  8.93 10.31 7.39 9.11 9.5 ...
##  $ station_avg_temp_c                   : num  26.4 26.9 26.8 25.8 26.6 ...
##  $ station_diur_temp_rng_c              : num  10.8 11.6 11.5 10.5 11.5 ...
##  $ station_max_temp_c                   : num  32.5 34 33 31.5 33.3 32 34 33 34 34 ...
##  $ station_min_temp_c                   : num  20.7 20.8 20.7 14.7 19.1 17 19.9 20.5 19 20 ...
##  $ station_precip_mm                    : num  3 55.6 38.1 30 4 11.5 72.9 50.1 89.2 78 ...
##  $ total_cases                          : int  0 0 0 0 0 0 0 0 0 0 ...
##  - attr(*, ".internal.selfref")=<externalptr>
##  - attr(*, "sorted")= chr  "city" "year" "weekofyear"
```

Only select data for 1 city: San Juan

```
sjdata <- trainprep[city=="sj"]
```

Splitting the data into training and testing datasets

Transform the Date column into the appropriate type

```
sjdata[,week_start_date := as.Date(week_start_date,format = "%Y-%m-%d")]
```

Look at date ranges

```
min(sjdata$week_start_date)
```

```
## [1] "1990-04-30"
```

```
max(sjdata$week_start_date)
```

```
## [1] "2008-04-22"
```

```
# Preserve 5 years of data for model validation
sjdata.train <- sjdata[week_start_date < "2003-01-01"]
sjdata.test <- sjdata[week_start_date >= "2003-01-01"]
```

Descriptive analytics and data visualisation

```
head(sjdata.train)
```

```
##    city year weekofyear week_start_date   ndvi_ne   ndvi_nw   ndvi_se
## 1:   sj 1990         18      1990-04-30 0.1226000 0.1037250 0.1984833
## 2:   sj 1990         19      1990-05-07 0.1699000 0.1421750 0.1623571
## 3:   sj 1990         20      1990-05-14 0.0322500 0.1729667 0.1572000
## 4:   sj 1990         21      1990-05-21 0.1286333 0.2450667 0.2275571
## 5:   sj 1990         22      1990-05-28 0.1962000 0.2622000 0.2512000
## 6:   sj 1990         23      1990-06-04        NA 0.1748500 0.2543143
##      ndvi_sw precipitation_amt_mm reanalysis_air_temp_k
## 1: 0.1776167                12.42              297.5729
## 2: 0.1554857                22.82              298.2114
## 3: 0.1708429                34.54              298.7814
## 4: 0.2358857                15.36              298.9871
## 5: 0.2473400                 7.52              299.5186
## 6: 0.1817429                 9.58              299.6300
##    reanalysis_avg_temp_k reanalysis_dew_point_temp_k
## 1:              297.7429                    292.4143
## 2:              298.4429                    293.9514
## 3:              298.8786                    295.4343
## 4:              299.2286                    295.3100
## 5:              299.6643                    295.8214
## 6:              299.7643                    295.8514
##    reanalysis_max_air_temp_k reanalysis_min_air_temp_k
## 1:                     299.8                     295.9
## 2:                     300.9                     296.4
## 3:                     300.5                     297.3
## 4:                     301.4                     297.0
## 5:                     301.9                     297.5
## 6:                     302.4                     298.1
##    reanalysis_precip_amt_kg_per_m2 reanalysis_relative_humidity_percent
## 1:                           32.00                              73.36571
## 2:                           17.94                              77.36857
## 3:                           26.10                              82.05286
## 4:                           13.90                              80.33714
## 5:                           12.20                              80.46000
## 6:                           26.49                              79.89143
##    reanalysis_specific_humidity_g_per_kg reanalysis_tdtr_k
## 1:                              14.01286          2.628571
## 2:                              15.37286          2.371429
## 3:                              16.84857          2.300000
```

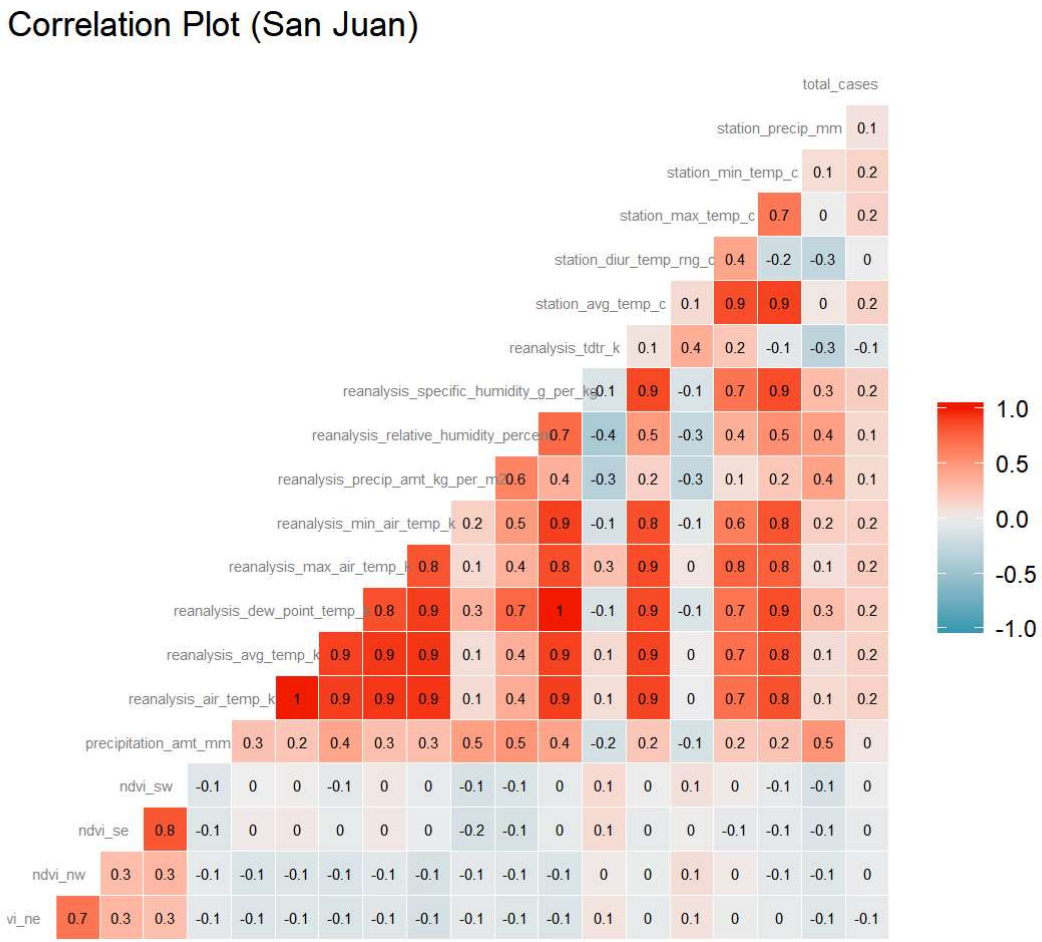
```
## 4:                            16.67286          2.428571
## 5:                            17.21000          3.014286
## 6:                            17.21286          2.100000
##    station_avg_temp_c station_diur_temp_rng_c station_max_temp_c
## 1:           25.44286                6.900000               29.4
## 2:           26.71429                6.371429               31.7
## 3:           26.71429                6.485714               32.2
## 4:           27.47143                6.771429               33.3
## 5:           28.94286                9.371429               35.0
## 6:           28.11429                6.942857               34.4
##    station_min_temp_c station_precip_mm total_cases
## 1:               20.0              16.0           4
## 2:               22.2               8.6           5
## 3:               22.8              41.4           4
## 4:               23.3               4.0           3
## 5:               23.9               5.8           6
## 6:               23.9              39.1           2
```

```
# Summary
summary(sjdata.train)
```

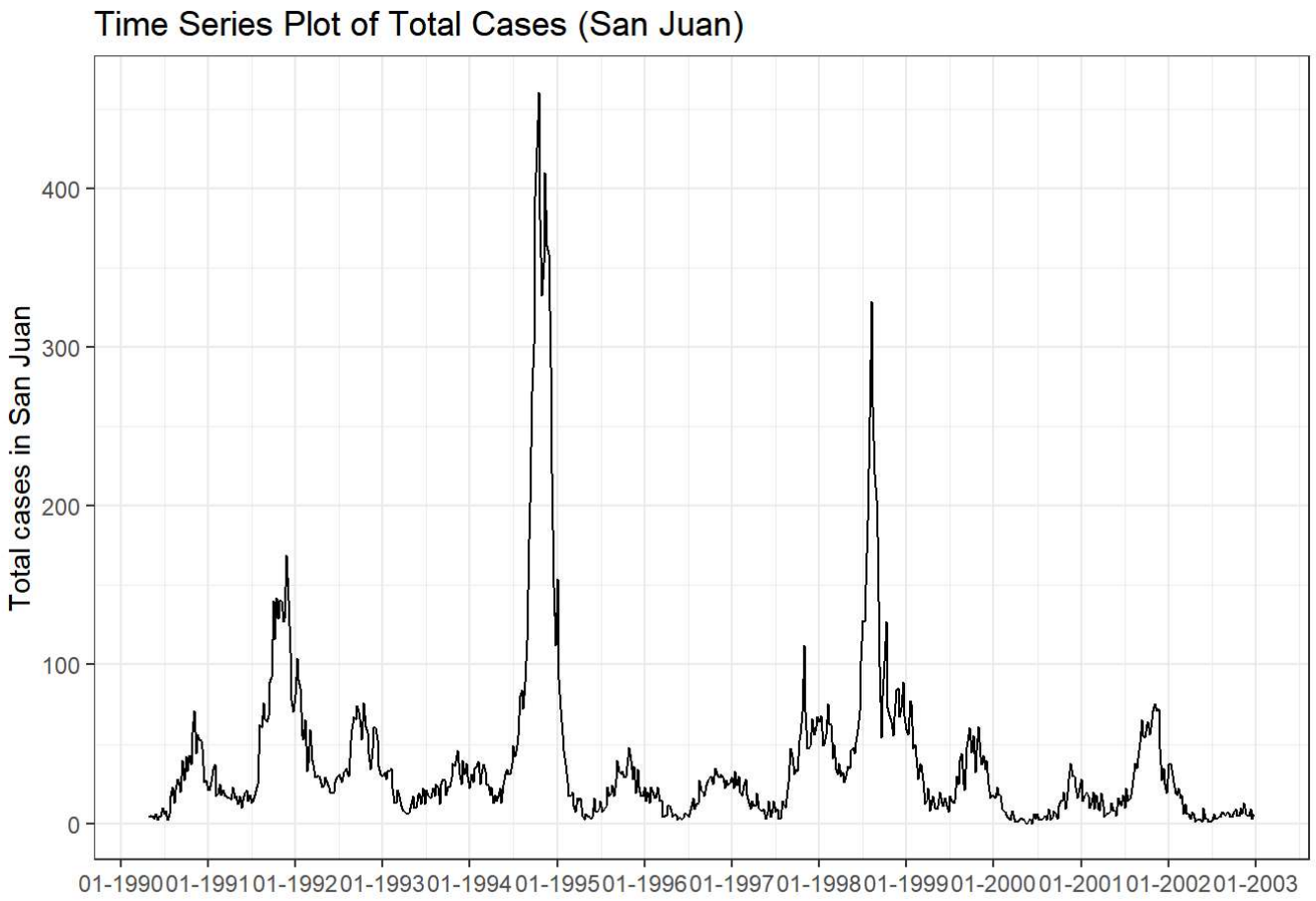
```
##      city                year        weekofyear    week_start_date
##  Length:659         Min.   :1990   Min.   : 1.00   Min.   :1990-04-30
##  Class :character   1st Qu.:1993   1st Qu.:14.00   1st Qu.:1993-06-28
##  Mode  :character   Median :1996   Median :27.00   Median :1996-08-26
##                     Mean   :1996   Mean   :26.95   Mean   :1996-08-26
##                     3rd Qu.:1999   3rd Qu.:40.00   3rd Qu.:1999-10-25
##                     Max.   :2002   Max.   :53.00   Max.   :2002-12-24
##
##     ndvi_ne            ndvi_nw            ndvi_se          ndvi_sw
##  Min.   :-0.29020   Min.   :-0.25280   Min.   :0.0360   Min.   :-0.06346
##  1st Qu.: 0.03332   1st Qu.: 0.04777   1st Qu.:0.1451   1st Qu.: 0.13303
##  Median : 0.08085   Median : 0.08838   Median :0.1806   Median : 0.16887
##  Mean   : 0.08504   Mean   : 0.09475   Mean   :0.1814   Mean   : 0.16859
##  3rd Qu.: 0.12755   3rd Qu.: 0.13681   3rd Qu.:0.2136   3rd Qu.: 0.20444
##  Max.   : 0.44627   Max.   : 0.43710   Max.   :0.3931   Max.   : 0.38142
##  NA's   :139        NA's   :39         NA's   :18       NA's   :18
##  precipitation_amt_mm reanalysis_air_temp_k reanalysis_avg_temp_k
##  Min.   :  0.0000     Min.   :295.9         Min.   :296.1
##  1st Qu.:  0.9975     1st Qu.:298.0         1st Qu.:298.2
##  Median : 21.1500     Median :299.2         Median :299.3
##  Mean   : 34.2270     Mean   :299.0         Mean   :299.1
##  3rd Qu.: 50.5800     3rd Qu.:300.0         3rd Qu.:300.1
##  Max.   :287.5500     Max.   :301.3         Max.   :301.4
##  NA's   :7            NA's   :4             NA's   :4
##  reanalysis_dew_point_temp_k reanalysis_max_air_temp_k
##  Min.   :289.6               Min.   :298.2
##  1st Qu.:293.8               1st Qu.:300.4
##  Median :295.4               Median :301.4
##  Mean   :295.1               Mean   :301.3
##  3rd Qu.:296.4               3rd Qu.:302.3
##  Max.   :297.5               Max.   :303.9
##  NA's   :4                   NA's   :4
##  reanalysis_min_air_temp_k reanalysis_precip_amt_kg_per_m2
##  Min.   :292.6             Min.   :  0.00
##  1st Qu.:296.2             1st Qu.: 11.75
##  Median :297.5             Median : 22.26
##  Mean   :297.2             Mean   : 31.99
##  3rd Qu.:298.2             3rd Qu.: 40.35
##  Max.   :299.5             Max.   :570.50
```

```
##  NA's   :4                NA's   :4
##  reanalysis_relative_humidity_percent
##  Min.   :66.74
##  1st Qu.:76.84
##  Median :79.16
##  Mean   :79.00
##  3rd Qu.:81.42
##  Max.   :87.30
##  NA's   :4
##  reanalysis_specific_humidity_g_per_kg reanalysis_tdtr_k
##  Min.   :11.72                         Min.   :1.357
##  1st Qu.:15.24                         1st Qu.:2.086
##  Median :16.83                         Median :2.357
##  Mean   :16.51                         Mean   :2.437
##  3rd Qu.:17.81                         3rd Qu.:2.693
##  Max.   :19.04                         Max.   :4.100
##  NA's   :4                             NA's   :4
##  station_avg_temp_c station_diur_temp_rng_c station_max_temp_c
##  Min.   :22.84      Min.   :4.529          Min.   :27.20
##  1st Qu.:25.93      1st Qu.:6.364          1st Qu.:30.60
##  Median :27.29      Median :6.857          Median :32.20
##  Mean   :27.05      Mean   :6.866          Mean   :31.72
##  3rd Qu.:28.19      3rd Qu.:7.386          3rd Qu.:32.80
##  Max.   :30.07      Max.   :9.914          Max.   :35.60
##  NA's   :4          NA's   :4              NA's   :4
##  station_min_temp_c station_precip_mm  total_cases
##  Min.   :17.80      Min.   :  0.00    Min.   :  0.00
##  1st Qu.:21.70      1st Qu.:  6.55    1st Qu.: 11.00
##  Median :22.80      Median : 16.60    Median : 23.00
##  Mean   :22.59      Mean   : 25.12    Mean   : 39.46
##  3rd Qu.:23.90      3rd Qu.: 32.40    3rd Qu.: 44.00
##  Max.   :25.60      Max.   :305.90    Max.   :461.00
##  NA's   :4          NA's   :4
```

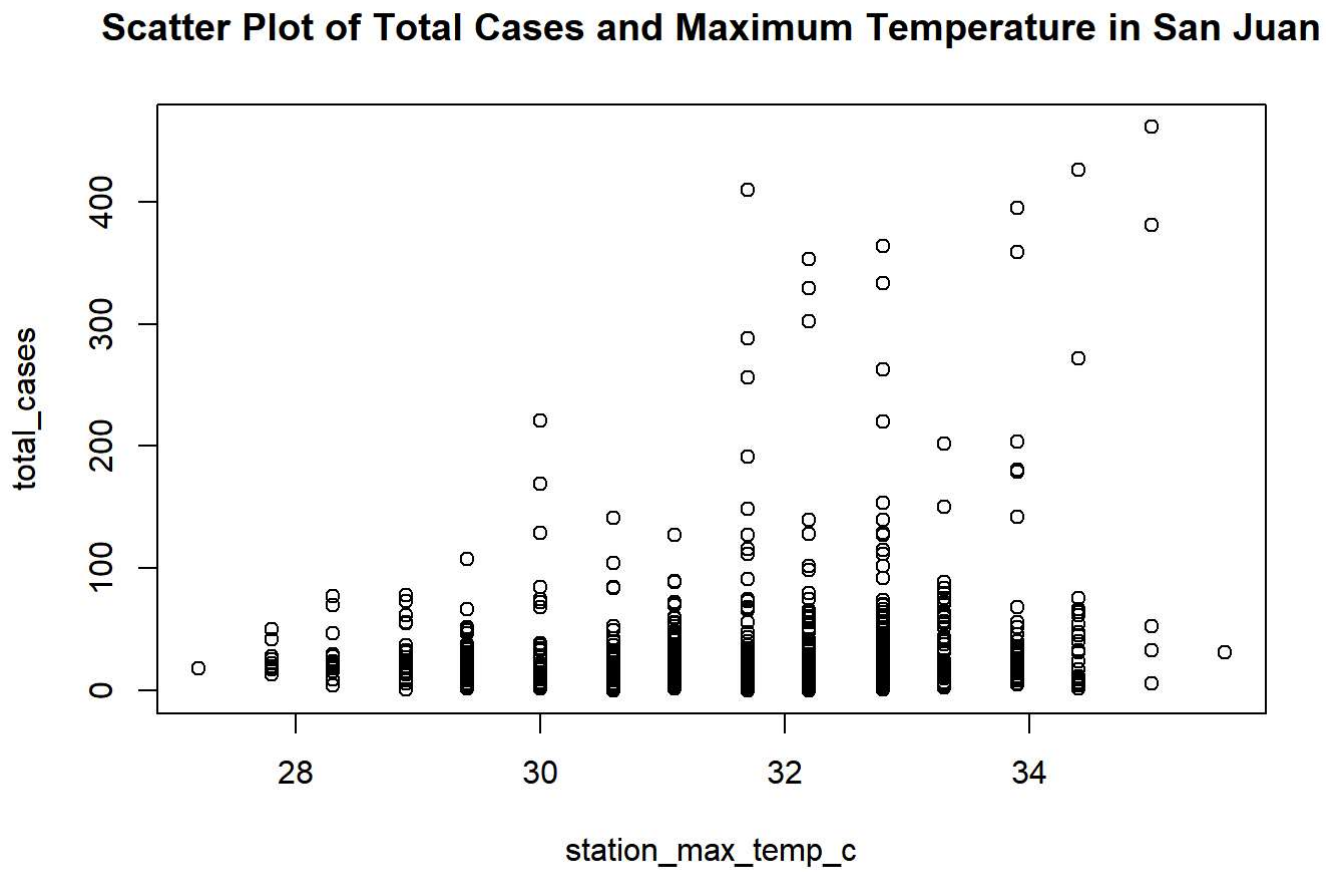
```
# Correlation plots
ggcorr(sjdata.train[, 5:24], label = TRUE, hjust = 0.85, size = 2,color = "grey50",
       label_size = 2) +
  ggplot2::labs(title = "Correlation Plot (San Juan)")
```



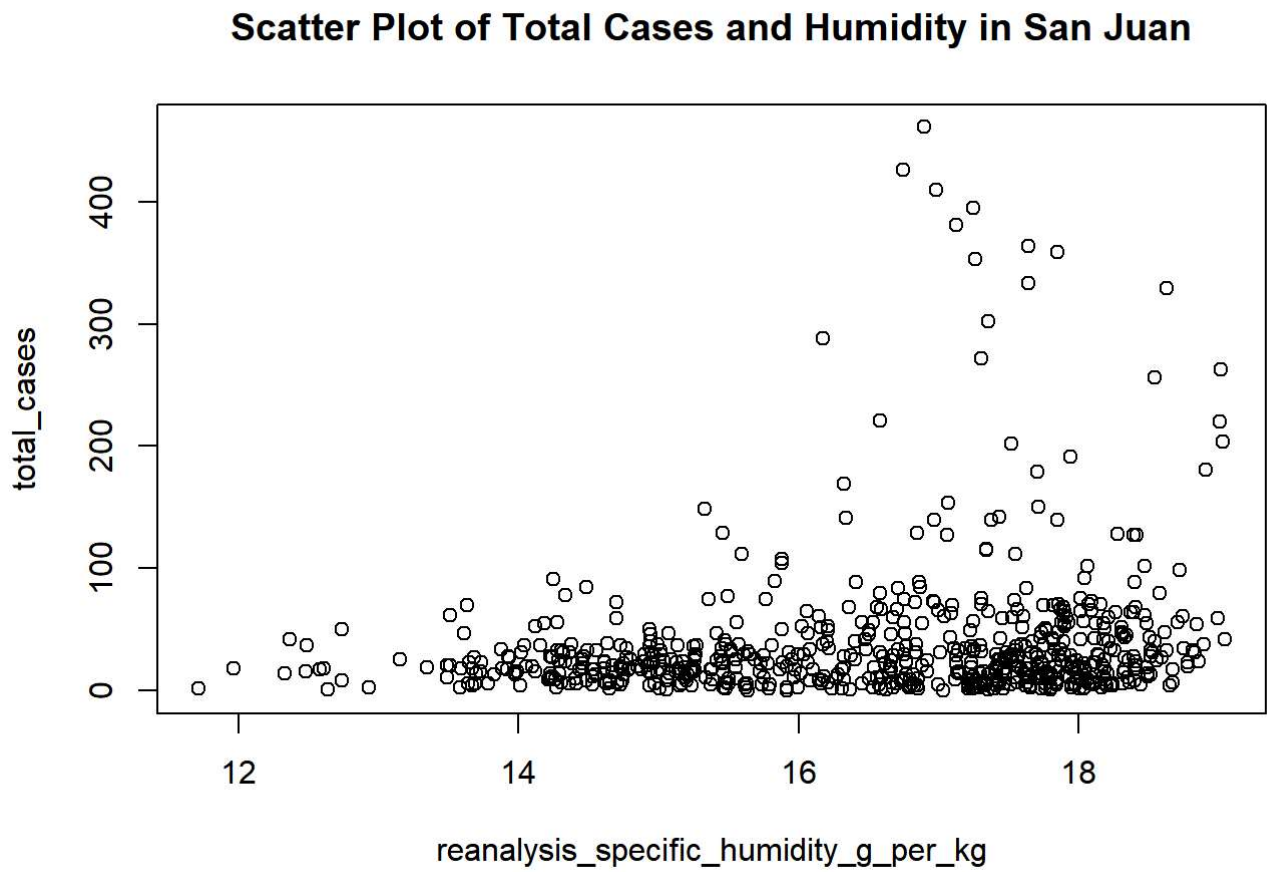

```
# Plot total cases with timeline -> seasonal effect
ggplot(sjdata.train , aes(week_start_date, total_cases)) +
  geom_line() +
  scale_y_continuous() +
  scale_x_date(breaks = date_breaks("year"), labels = date_format("%m-%Y")) +
  ylab("Total cases in San Juan") +
  xlab("")+
  theme(axis.text.x=element_text(angle=-70, hjust=0.001)) +
  labs(title = "Time Series Plot of Total Cases (San Juan)") +
  theme_bw()
```



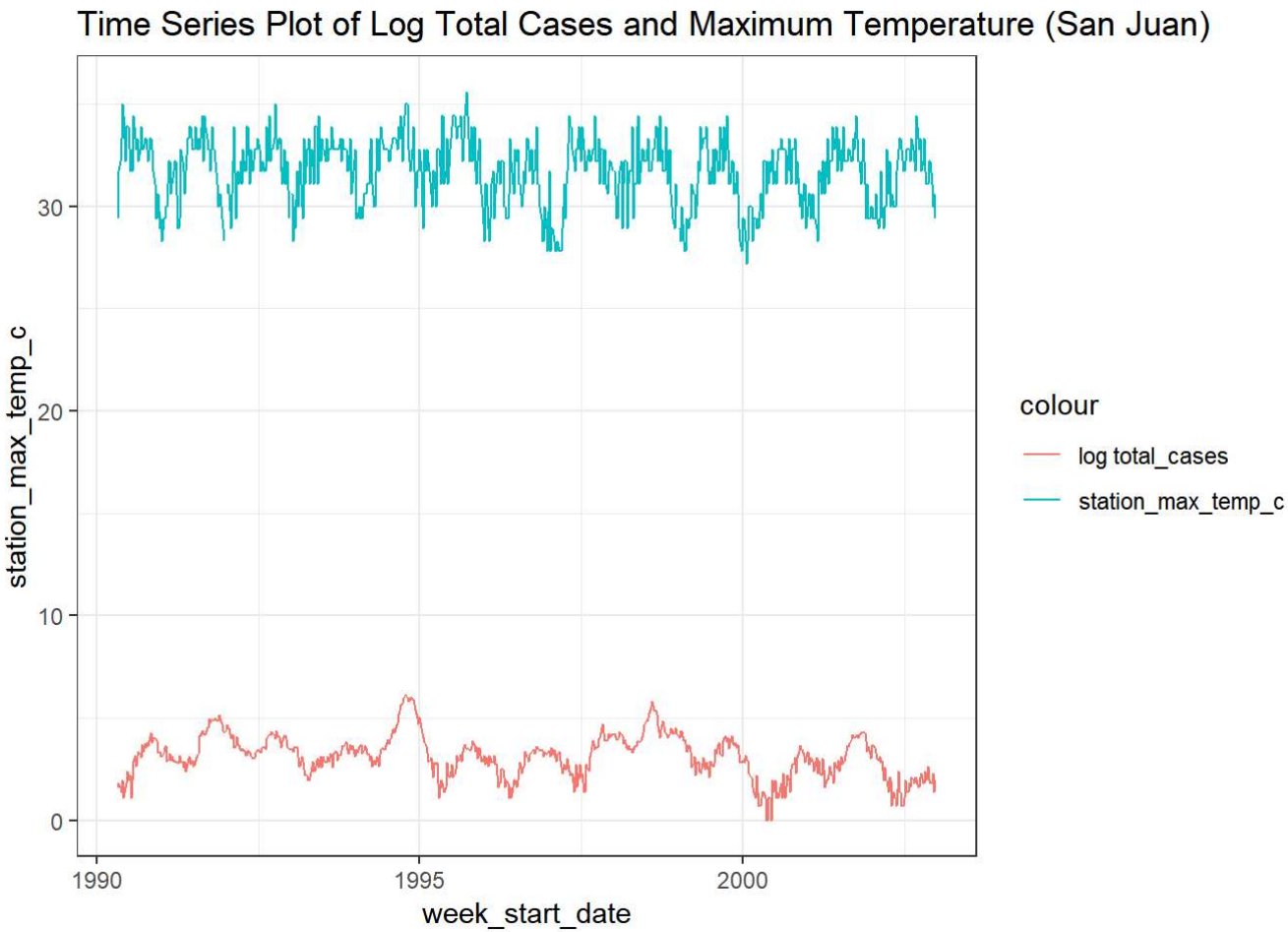

```
# Scatter Plot Total cases vs. Max Temperature
plot(total_cases ~ station_max_temp_c, data=sjdata.train
     ,main = "Scatter Plot of Total Cases and Maximum Temperature in San Juan")
```



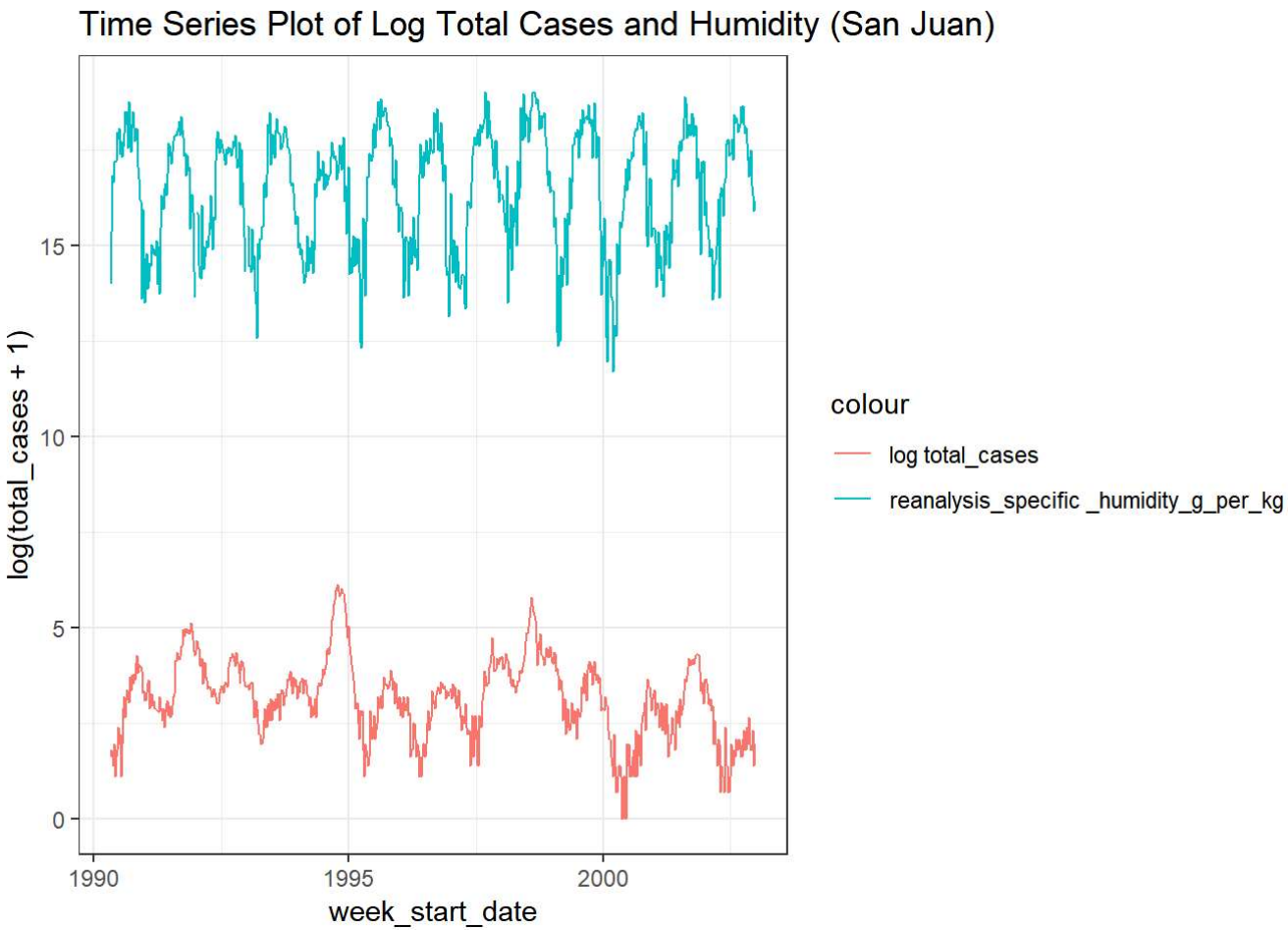

```
# Scatter Plot Total cases vs. Humidity
plot(total_cases ~ reanalysis_specific_humidity_g_per_kg, data=sjdata.train,
     main = "Scatter Plot of Total Cases and Humidity in San Juan")
```




```
# Time Series Plot of log(Total cases) & Max Temperature
ggplot(sjdata.train, aes(week_start_date)) +
  geom_line(aes(y = station_max_temp_c, colour = "station_max_temp_c")) +
  geom_line(aes(y = (y = log(total_cases+1)),
                colour = "log total_cases")) +
  labs(title = "Time Series Plot of Log Total Cases and Maximum Temperature (San Juan)") +
  theme_bw()
```




```
# Time Series Plot of log(Total cases) & Humidity
ggplot(sjdata.train, aes(week_start_date)) +
  geom_line(aes(y = log(total_cases+1),
                colour = "log total_cases")) +
  geom_line(aes(y = reanalysis_specific_humidity_g_per_kg,
                colour = "reanalysis_specific _humidity_g_per_kg")) +
  labs(title = "Time Series Plot of Log Total Cases and Humidity (San Juan)") +
  theme_bw()
```



Data imputation

```
# Build a function to impute data with the most recent that is non missing (Using LOCF)
na.locf.data.frame <-
  function(object, ...) replace(object, TRUE, lapply(object, na.locf, ...))

# Fill in NAs
sjdata.train.imputed <- na.locf.data.frame(sjdata.train)
summary(sjdata.train.imputed)
```

```
##      city                year        weekofyear    week_start_date
##  Length:659         Min.   :1990   Min.   : 1.00   Min.   :1990-04-30
##  Class :character   1st Qu.:1993   1st Qu.:14.00   1st Qu.:1993-06-28
##  Mode  :character   Median :1996   Median :27.00   Median :1996-08-26
##                     Mean   :1996   Mean   :26.95   Mean   :1996-08-26
##                     3rd Qu.:1999   3rd Qu.:40.00   3rd Qu.:1999-10-25
##                     Max.   :2002   Max.   :53.00   Max.   :2002-12-24
##     ndvi_ne            ndvi_nw            ndvi_se          ndvi_sw
##  Min.   :-0.29020   Min.   :-0.25280   Min.   :0.0360   Min.   :-0.06346
##  1st Qu.: 0.03343   1st Qu.: 0.04863   1st Qu.:0.1405   1st Qu.: 0.13411
##  Median : 0.07740   Median : 0.08460   Median :0.1787   Median : 0.16953
##  Mean   : 0.08379   Mean   : 0.09341   Mean   :0.1792   Mean   : 0.16885
##  3rd Qu.: 0.12387   3rd Qu.: 0.13375   3rd Qu.:0.2127   3rd Qu.: 0.20316
##  Max.   : 0.44627   Max.   : 0.43710   Max.   :0.3931   Max.   : 0.38142
##  precipitation_amt_mm reanalysis_air_temp_k reanalysis_avg_temp_k
##  Min.   :  0.000      Min.   :295.9         Min.   :296.1
##  1st Qu.:  0.245      1st Qu.:298.0         1st Qu.:298.2
##  Median : 20.800      Median :299.2         Median :299.3
##  Mean   : 33.962      Mean   :299.0         Mean   :299.1
##  3rd Qu.: 50.025      3rd Qu.:300.0         3rd Qu.:300.1
##  Max.   :287.550      Max.   :301.3         Max.   :301.4
##  reanalysis_dew_point_temp_k reanalysis_max_air_temp_k
##  Min.   :289.6               Min.   :298.2
##  1st Qu.:293.8               1st Qu.:300.4
##  Median :295.4               Median :301.4
##  Mean   :295.1               Mean   :301.3
##  3rd Qu.:296.4               3rd Qu.:302.2
##  Max.   :297.5               Max.   :303.9
##  reanalysis_min_air_temp_k reanalysis_precip_amt_kg_per_m2
##  Min.   :292.6             Min.   :  0.00
##  1st Qu.:296.2             1st Qu.: 11.85
##  Median :297.5             Median : 22.10
##  Mean   :297.2             Mean   : 31.92
##  3rd Qu.:298.2             3rd Qu.: 40.30
##  Max.   :299.5             Max.   :570.50
##  reanalysis_relative_humidity_percent
##  Min.   :66.74
##  1st Qu.:76.80
##  Median :79.15
```

```
##  Mean   :78.99
##  3rd Qu.:81.41
##  Max.   :87.30
##  reanalysis_specific_humidity_g_per_kg reanalysis_tdtr_k
##  Min.   :11.72                         Min.   :1.357
##  1st Qu.:15.23                         1st Qu.:2.086
##  Median :16.82                         Median :2.357
##  Mean   :16.50                         Mean   :2.436
##  3rd Qu.:17.80                         3rd Qu.:2.686
##  Max.   :19.04                         Max.   :4.100
##  station_avg_temp_c station_diur_temp_rng_c station_max_temp_c
##  Min.   :22.84      Min.   :4.529           Min.   :27.20
##  1st Qu.:25.92      1st Qu.:6.357           1st Qu.:30.60
##  Median :27.27      Median :6.843           Median :32.20
##  Mean   :27.04      Mean   :6.863           Mean   :31.71
##  3rd Qu.:28.19      3rd Qu.:7.386           3rd Qu.:32.80
##  Max.   :30.07      Max.   :9.914           Max.   :35.60
##  station_min_temp_c station_precip_mm  total_cases
##  Min.   :17.80      Min.   :  0.00    Min.   :  0.00
##  1st Qu.:21.70      1st Qu.:  6.60    1st Qu.: 11.00
##  Median :22.80      Median : 16.60    Median : 23.00
##  Mean   :22.58      Mean   : 25.09    Mean   : 39.46
##  3rd Qu.:23.90      3rd Qu.: 32.40    3rd Qu.: 44.00
##  Max.   :25.60      Max.   :305.90    Max.   :461.00
```

Diagnostic study

```
# Convert data to TS with weekly frequency
sjtrain.ts <- ts(sjdata.train$total_cases, frequency = 52, start = c(1990,18))
head(sjtrain.ts)
```

```
## Time Series:
## Start = c(1990, 18)
## End = c(1990, 23)
## Frequency = 52
## [1] 4 5 4 3 6 2
```

Stationarity test

```
# ADF Test
adf.test(sjdata.train$total_cases)
```

```
## Warning in adf.test(sjdata.train$total_cases): p-value smaller than printed
## p-value
```

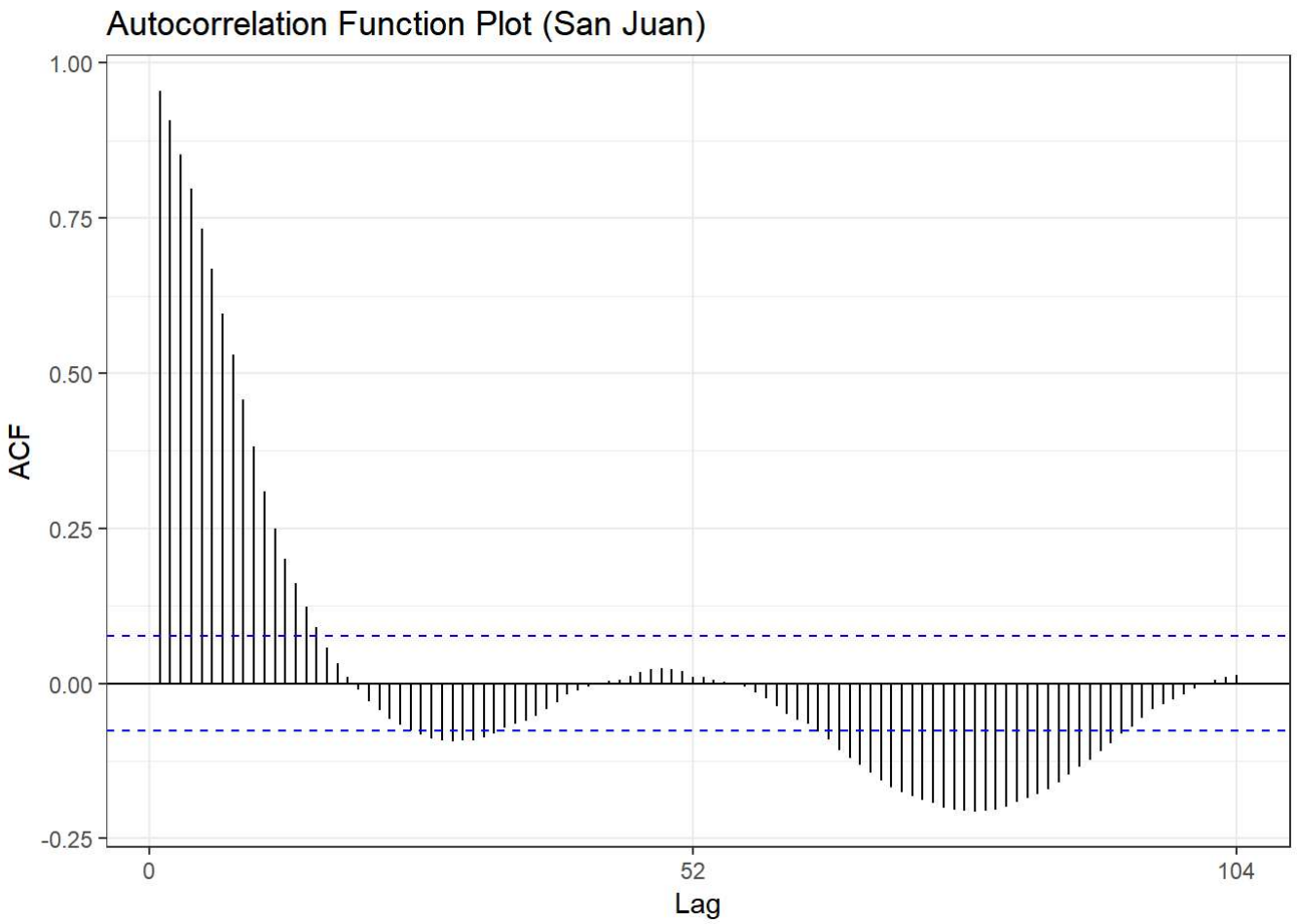
```
##
##  Augmented Dickey-Fuller Test
##
## data:  sjdata.train$total_cases
## Dickey-Fuller = -5.8915, Lag order = 8, p-value = 0.01
## alternative hypothesis: stationary
```

```
# KPSS Test
kpss.test(sjdata.train$total_cases)
```

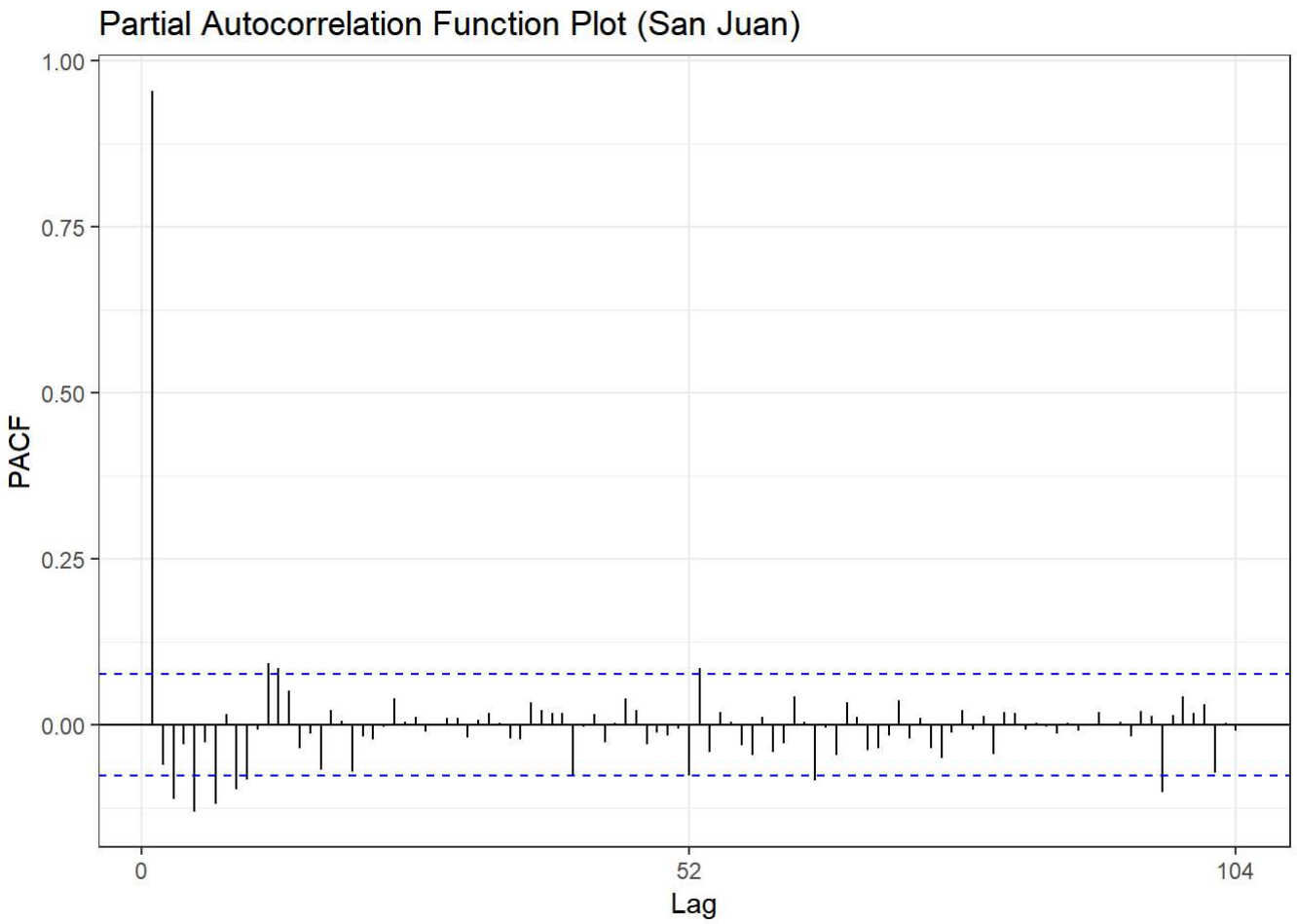
```
##
##  KPSS Test for Level Stationarity
##
## data:  sjdata.train$total_cases
## KPSS Level = 0.36839, Truncation lag parameter = 6, p-value =
## 0.09078
```

The time series is stationary.

```
# ACF Plot
ggAcf(sjtrain.ts) +
  labs(title = "Autocorrelation Function Plot (San Juan)") +
  theme_bw()
```



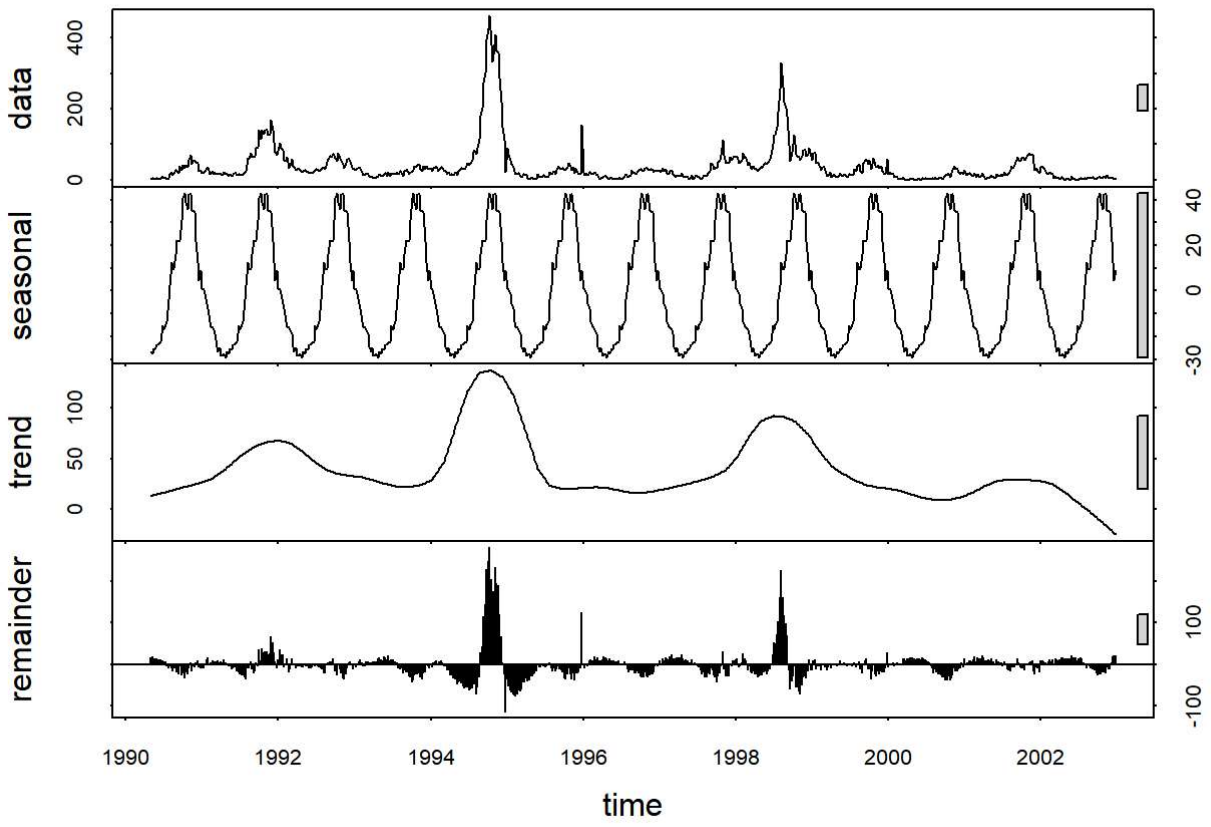

```
# PACF Plot
ggPacf(sjtrain.ts)  +
  labs(title = "Partial Autocorrelation Function Plot (San Juan)") +
  theme_bw()
```



We can observe significant autocorrelation and seasonality.

Decompose the time series

```
decomp = stl(sjtrain.ts, s.window="periodic")
plot(decomp)
```



Modeling

The following models were used for forecasting dengue spread.

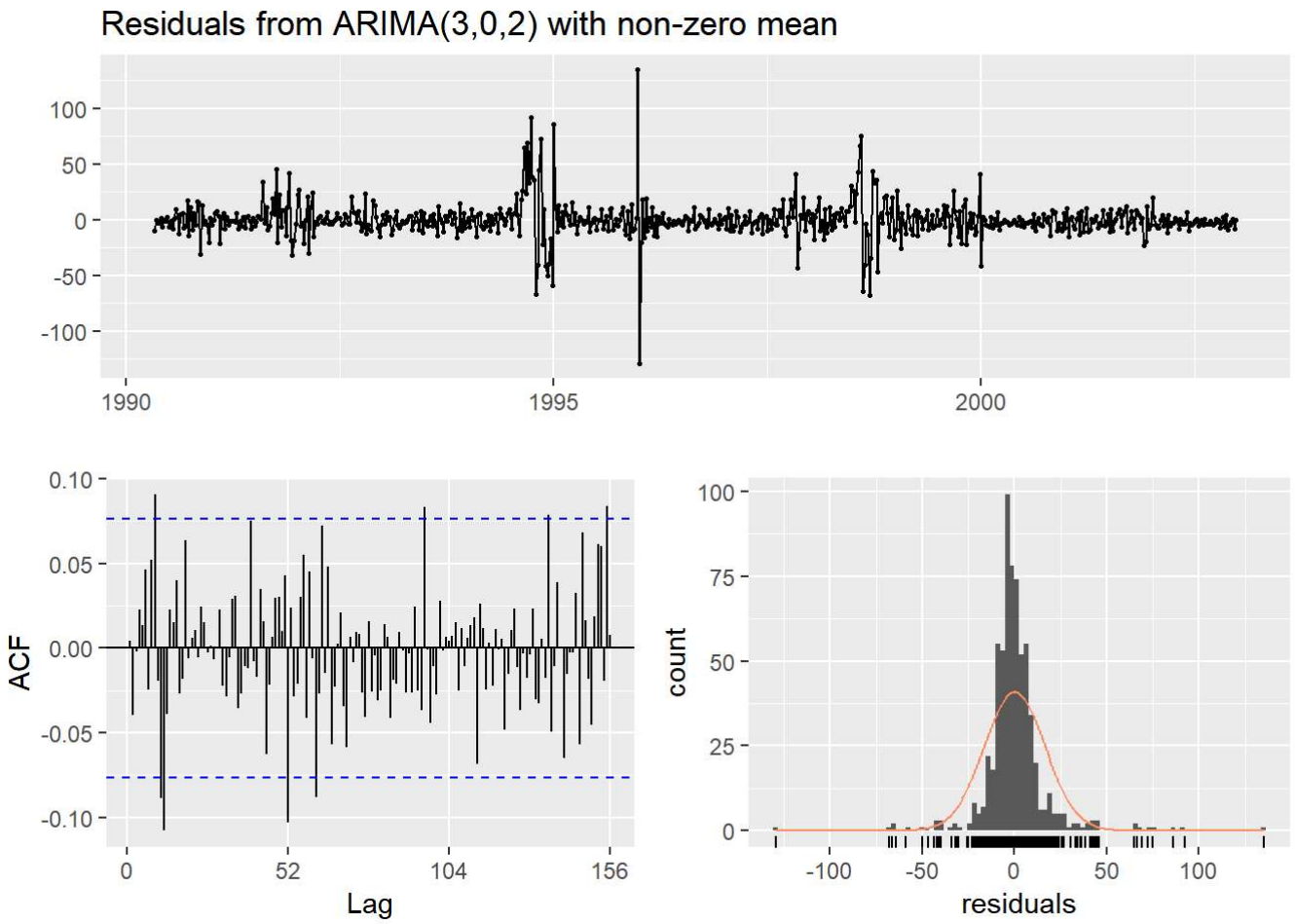
ARIMA

Use ARIMA without seasonality on the training data and analyse the residuals.

```
sj.fit1 <- auto.arima(sjtrain.ts, seasonal = F)
sj.fit1
```

```
## Series: sjtrain.ts
## ARIMA(3,0,2) with non-zero mean
##
## Coefficients:
##          ar1     ar2      ar3     ma1      ma2     mean
##       0.9203  0.8697  -0.8167  0.0726  -0.7310  39.0775
## s.e.  0.0598  0.0761   0.0542  0.0741   0.0706   8.2120
##
## sigma^2 estimated as 276:  log likelihood=-2785.37
## AIC=5584.74   AICc=5584.91   BIC=5616.17
```

```
checkresiduals(sj.fit1)
```



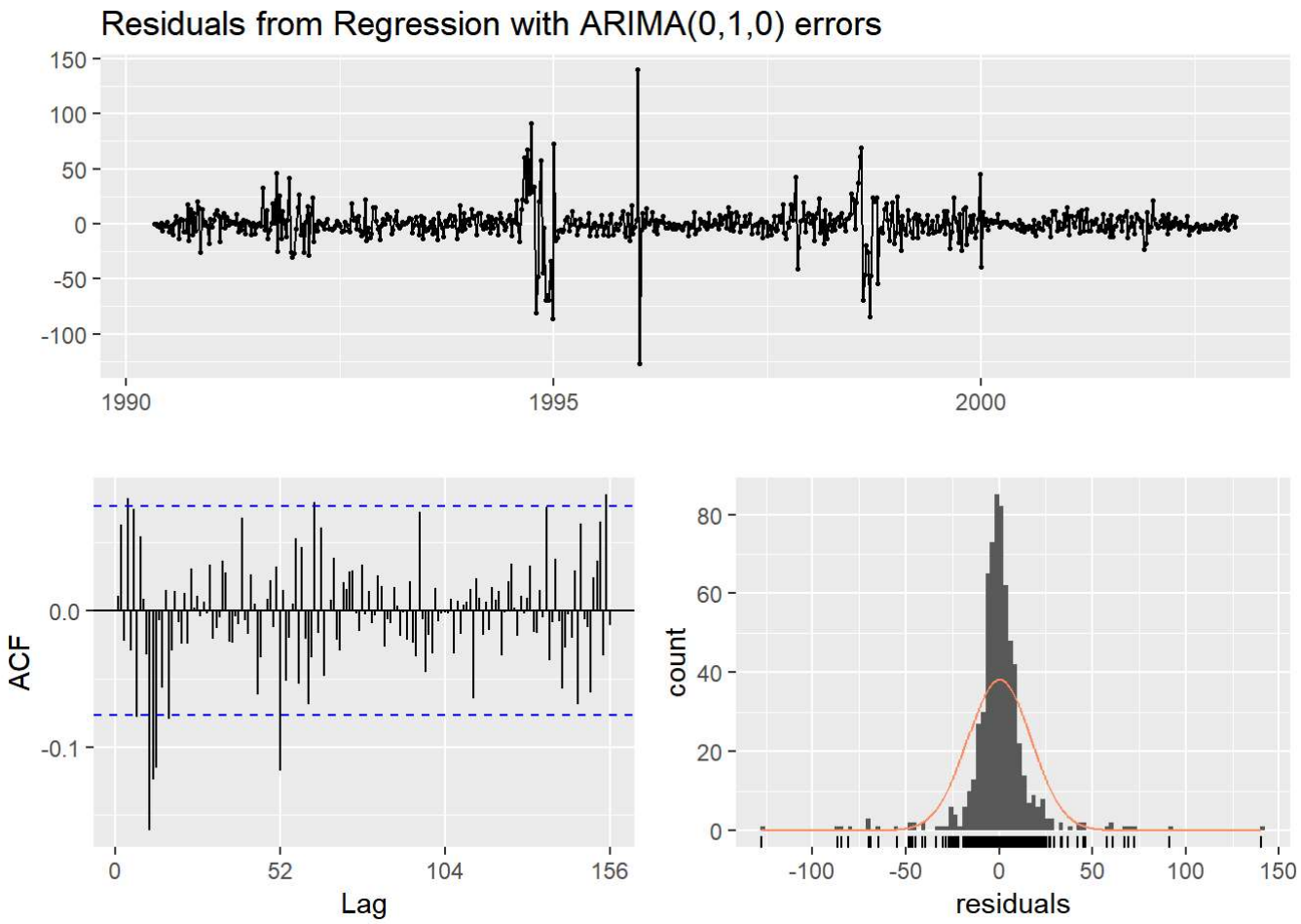

```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(3,0,2) with non-zero mean
## Q* = 96.188, df = 98, p-value = 0.5329
##
## Model df: 6.   Total lags used: 104
```

Let's take the fourier transform of the signal and evaluate the model residuals

```
fit <- list(aicc=Inf)
for(i in 1:25)
{
  fit1 <- auto.arima(sjtrain.ts, xreg=fourier(sjtrain.ts, K=i), seasonal=F)
  if(fit$aicc < fit1$aicc)
    fit1 <- fit
  else break;
}
fit1
```

```
## Series: sjtrain.ts
## Regression with ARIMA(0,1,0) errors
##
## Coefficients:
##        S1-52     C1-52
##       0.3589  -32.6728
## s.e.  7.7111    7.7396
##
## sigma^2 estimated as 287.2:  log likelihood=-2794.82
## AIC=5595.64   AICc=5595.68   BIC=5609.11
```

```
checkresiduals(fit1)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from Regression with ARIMA(0,1,0) errors
## Q* = 124.7, df = 102, p-value = 0.06292
##
## Model df: 2.   Total lags used: 104
```

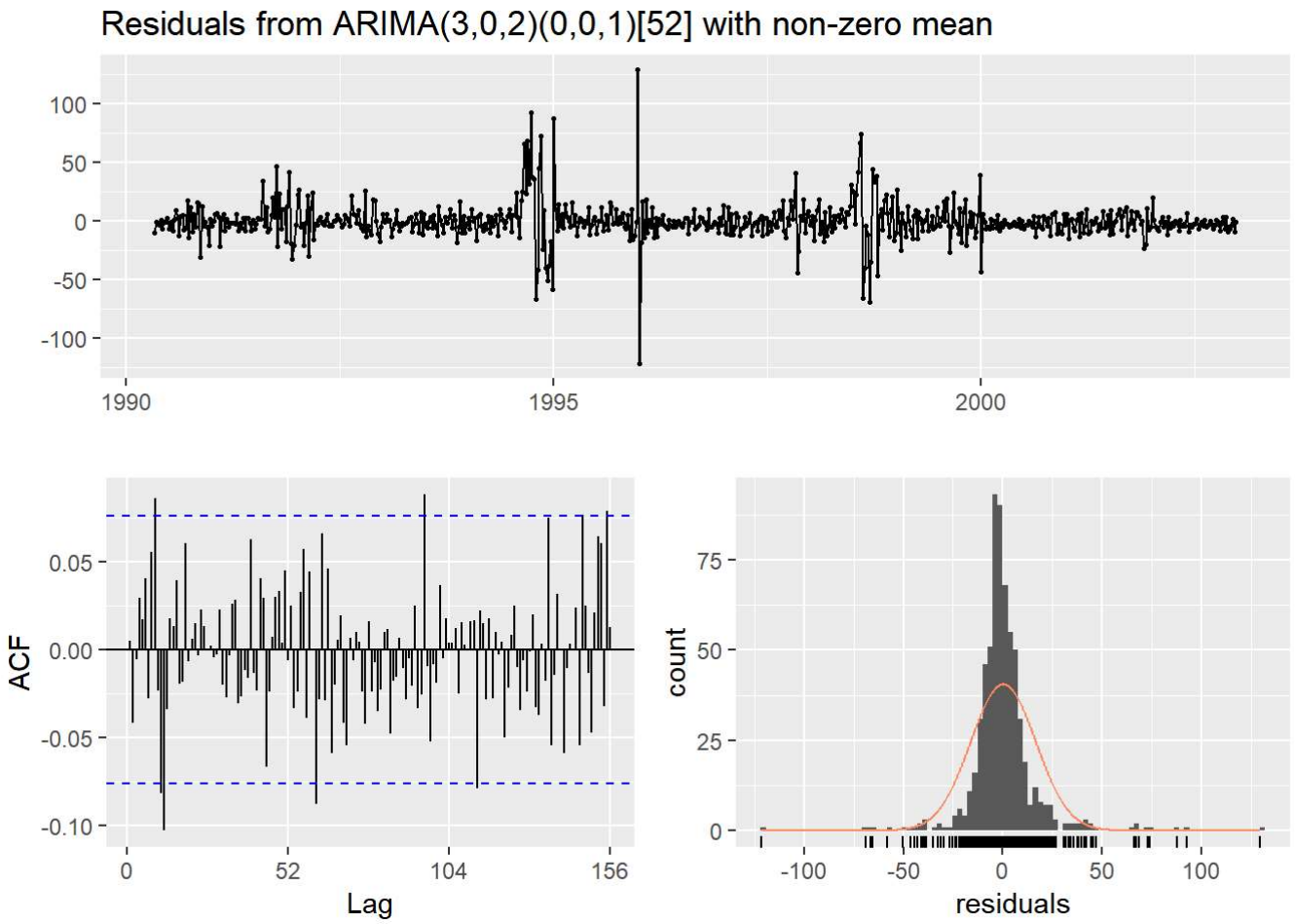
SARIMA

Apply Seasonal ARIMA to the data and analyse the residuals.

```
sj.fit2 <- auto.arima(sjtrain.ts, seasonal = T)
sj.fit2
```

```
## Series: sjtrain.ts
## ARIMA(3,0,2)(0,0,1)[52] with non-zero mean
##
## Coefficients:
##          ar1     ar2      ar3     ma1      ma2     sma1     mean
##       0.9216  0.8803  -0.8287  0.0773  -0.7426  -0.1002  39.3384
## s.e.  0.0566  0.0749   0.0518  0.0710   0.0682   0.0382   7.2421
##
## sigma^2 estimated as 273.4:  log likelihood=-2781.97
## AIC=5579.94   AICc=5580.16   BIC=5615.86
```

```
checkresiduals(sj.fit2)
```



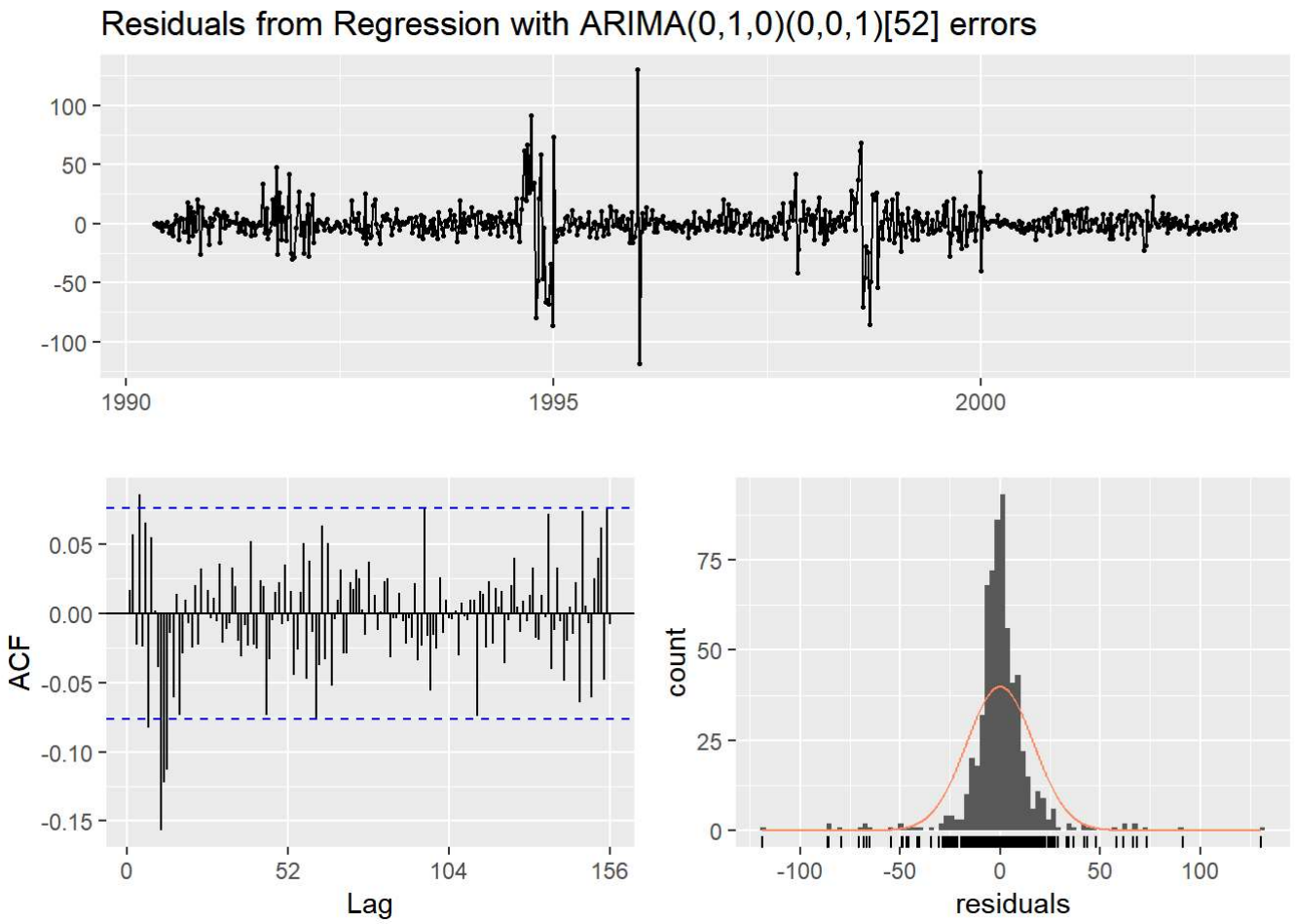

```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(3,0,2)(0,0,1)[52] with non-zero mean
## Q* = 88.348, df = 97, p-value = 0.7232
##
## Model df: 7.   Total lags used: 104
```

Let's take the fourier transform of the signal and evaluate the model residuals

```
fit <- list(aicc=Inf)
for(i in 1:25)
{
  fit1 <- auto.arima(sjtrain.ts, xreg=fourier(sjtrain.ts, K=i), seasonal=T)
  if(fit$aicc < fit1$aicc)
    fit1 <- fit
  else break;
}
(fit1)
```

```
## Series: sjtrain.ts
## Regression with ARIMA(0,1,0)(0,0,1)[52] errors
##
## Coefficients:
##          sma1   S1-52     C1-52
##       -0.1151  0.4835  -33.0620
## s.e.   0.0383  6.8444    6.8712
##
## sigma^2 estimated as 283.4:  log likelihood=-2790.36
## AIC=5588.73   AICc=5588.79   BIC=5606.68
```

```
checkresiduals(fit1)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from Regression with ARIMA(0,1,0)(0,0,1)[52] errors
## Q* = 114.04, df = 101, p-value = 0.1769
##
## Model df: 3.   Total lags used: 104
```

SARIMAX

```
# prepare xreg
sjtrain.ts1 <- ts(sjdata.train.imputed,
                  freq=365.25/7,
                  start=decimal_date(ymd("1990-04-30")))
```

```
## Warning in data.matrix(data): NAs introduced by coercion
```

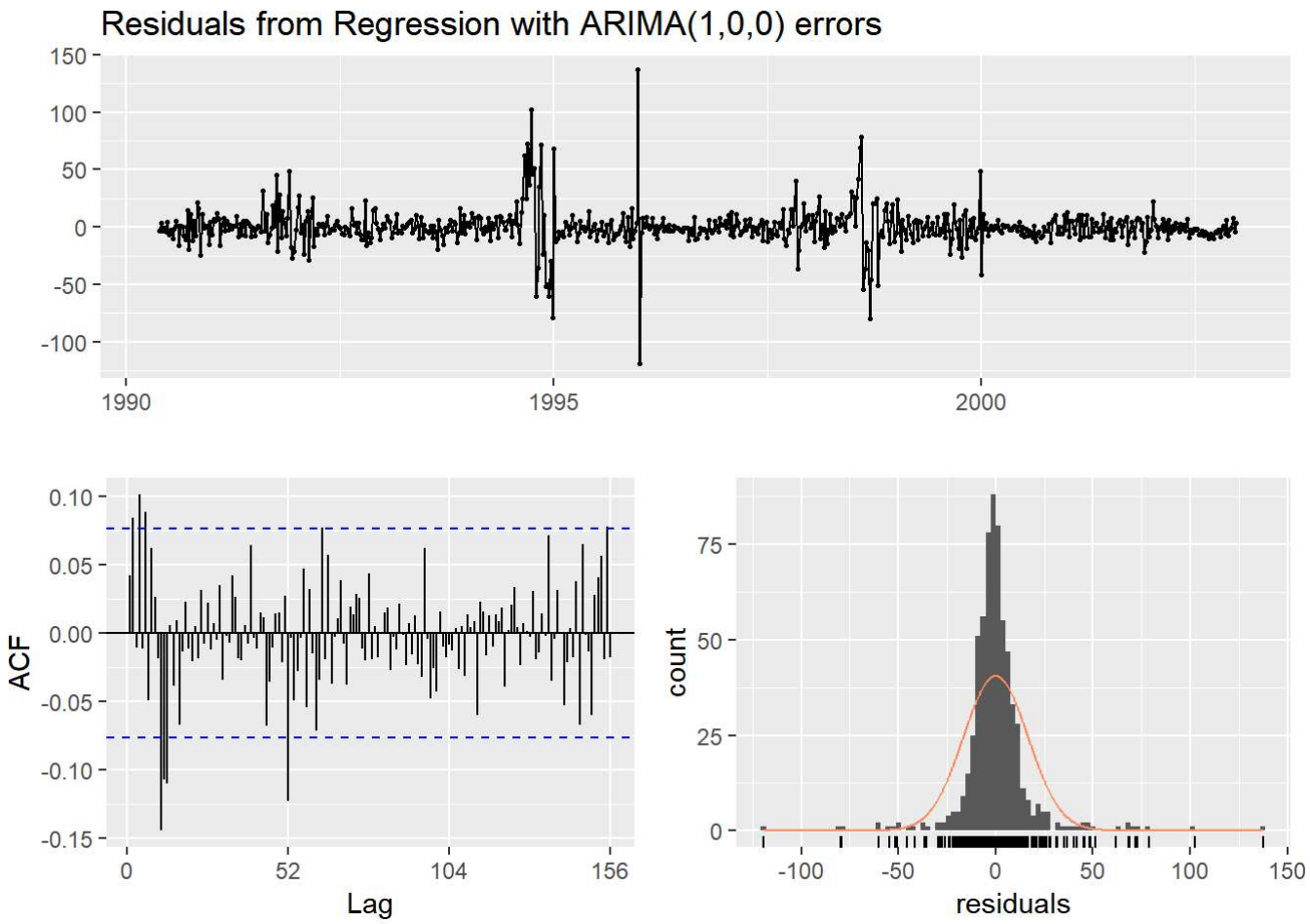
```
varlist <- c("precipitation_amt_mm"
             ,"reanalysis_dew_point_temp_k"
             ,"reanalysis_relative_humidity_percent"
             ,"station_avg_temp_c"
             ,"station_diur_temp_rng_c"
             ,"station_max_temp_c"
             ,"station_min_temp_c")

create.tslag <- function(x, dataset) {
  cbind(
    Lag0 = dataset[,x],
    Lag1 = stats::lag(dataset[,x],-1),
    Lag2 = stats::lag(dataset[,x],-2),
    Lag3 = stats::lag(dataset[,x],-3),
    Lag4 = stats::lag(dataset[,x],-4)) %>%
    head(NROW(dataset))
}
```

```
precipitation <- create.tslag(x = "precipitation_amt_mm"
                              ,sjtrain.ts1)

dew.temp <- create.tslag(x = "reanalysis_dew_point_temp_k"
                         ,sjtrain.ts1)

relative.humidity <- create.tslag(x = "reanalysis_relative_humidity_percent"
                                  ,sjtrain.ts1)

avg.temp <- create.tslag(x = "station_avg_temp_c"
                         ,sjtrain.ts1)

diur.temp <- create.tslag(x = "station_diur_temp_rng_c"
                          ,sjtrain.ts1)

max.temp <- create.tslag(x = "station_max_temp_c"
                         ,sjtrain.ts1)

min.temp <- create.tslag(x = "station_min_temp_c"
                         ,sjtrain.ts1)
```

```
# CCF plot -> week 3,4
#differencing
sjtrain.ts.diff1 = diff(sjtrain.ts,1)
dew.temp.diff1 = diff(dew.temp[,1],1)
#ccf(dew.temp.diff1,
#    sjtrain.ts.diff1,
#    type = c("correlation","covariance"))

# final model
fw <- fourier(sjtrain.ts, K=2)

fit2 <- auto.arima(sjtrain.ts, xreg=cbind(precipitation[,4], relative.humidity[,4]
                                          ,dew.temp[,4], avg.temp[,4]
                                          ,diur.temp[,4],max.temp[,4]
                                          ,min.temp[,4],fw))

fit2
```

```
## Series: sjtrain.ts
## Regression with ARIMA(1,0,0) errors
##
## Coefficients:
##          ar1  precipitation[, 4]  relative.humidity[, 4]  dew.temp[, 4]
##       0.9493              0.0008                 -0.3712         0.1521
## s.e.  0.0120              0.0146                  0.2453         0.1239
##       avg.temp[, 4]  diur.temp[, 4]  max.temp[, 4]  min.temp[, 4]
##              1.1380          0.7611        -0.4555         0.0222
## s.e.         1.8338          1.0520         0.7786         0.9886
##       fw.S1-52  fw.C1-52  fw.S2-52  fw.C2-52
##          0.154  -31.9520   -1.4517    5.8086
## s.e.     7.298    7.2398    3.8414    3.8215
##
## sigma^2 estimated as 280.6:  log likelihood=-2776.37
## AIC=5578.73   AICc=5579.3   BIC=5637.05
```

```
checkresiduals(fit2)
```

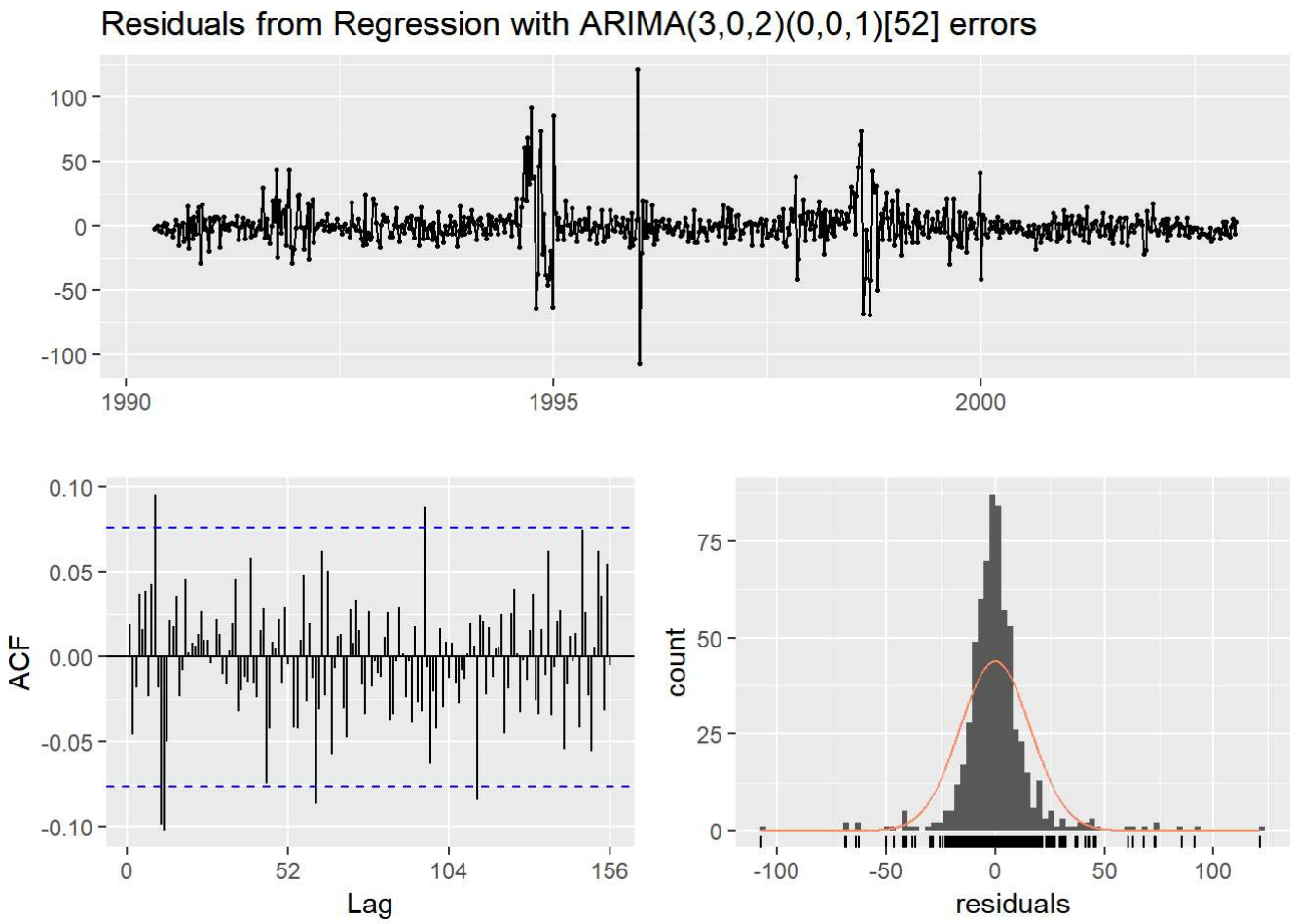



```
##
##  Ljung-Box test
##
## data:  Residuals from Regression with ARIMA(1,0,0) errors
## Q* = 118.96, df = 92, p-value = 0.03082
##
## Model df: 12.   Total lags used: 104
```

```
fit3 <- auto.arima(sjtrain.ts, xreg=cbind(precipitation[,1],relative.humidity[,1]
                                          ,dew.temp[,1],avg.temp[,1]
                                          ,diur.temp[,1],max.temp[,1]
                                          ,min.temp[,1],fw))
fit3
```

```
## Series: sjtrain.ts
## Regression with ARIMA(3,0,2)(0,0,1)[52] errors
##
## Coefficients:
##          ar1     ar2      ar3     ma1      ma2     sma1
##       0.8374  0.8970  -0.7678  0.1474  -0.6931  -0.1187
## s.e.  0.0817  0.0572   0.0738  0.0953   0.0909   0.0393
##       precipitation[, 1]  relative.humidity[, 1]  dew.temp[, 1]
##                  -0.0041                  0.0305         -0.165
## s.e.              0.0137                  0.2364          0.111
##       avg.temp[, 1]  diur.temp[, 1]  max.temp[, 1]  min.temp[, 1]
##              4.2904          0.5500        -0.0037        -1.5107
## s.e.         1.6973          0.9813         0.7191         0.9241
##       fw.S1-52  fw.C1-52  fw.S2-52  fw.C2-52
##        -3.3772  -31.2108   -1.6327    5.4537
## s.e.    8.5763    8.4247    4.2137    4.1986
##
## sigma^2 estimated as 265.4:  log likelihood=-2767.12
## AIC=5570.23   AICc=5571.3   BIC=5651.06
```

```
checkresiduals(fit3)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from Regression with ARIMA(3,0,2)(0,0,1)[52] errors
## Q* = 88.586, df = 87, p-value = 0.4325
##
## Model df: 17.   Total lags used: 104
```

Neural Network

```
# Fit Neural Network
fit4 <- nnetar(sjtrain.ts,xreg=sjtrain.ts1[,5:23])
fit4
```

```
## 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 7.896
```

Model Selection

```
# Fill in NAs
sjdata.test.imputed <- na.locf.data.frame(sjdata.test)
summary(sjdata.test.imputed)
```

```
##      city                year        weekofyear    week_start_date
##  Length:277         Min.   :2003   Min.   : 1.00   Min.   :2003-01-01
##  Class :character   1st Qu.:2004   1st Qu.:12.00   1st Qu.:2004-04-29
##  Mode  :character   Median :2005   Median :25.00   Median :2005-08-27
##                     Mean   :2005   Mean   :25.43   Mean   :2005-08-26
##                     3rd Qu.:2006   3rd Qu.:39.00   3rd Qu.:2006-12-24
##                     Max.   :2008   Max.   :53.00   Max.   :2008-04-22
##     ndvi_ne            ndvi_nw             ndvi_se
##  Min.   :-0.40625   Min.   :-0.456100   Min.   :-0.01553
##  1st Qu.:-0.05557   1st Qu.:-0.042350   1st Qu.: 0.12332
##  Median : 0.01010   Median : 0.010950   Median : 0.16833
##  Mean   :-0.00306   Mean   : 0.004001   Mean   : 0.16937
##  3rd Qu.: 0.05160   3rd Qu.: 0.053400   3rd Qu.: 0.20832
##  Max.   : 0.49340   Max.   : 0.296000   Max.   : 0.35434
##     ndvi_sw        precipitation_amt_mm reanalysis_air_temp_k
##  Min.   :0.01025   Min.   :  0.00       Min.   :296.1
##  1st Qu.:0.12143   1st Qu.:  0.00       1st Qu.:298.4
##  Median :0.15679   Median : 20.30       Median :299.5
##  Mean   :0.16033   Mean   : 38.15       Mean   :299.5
##  3rd Qu.:0.19631   3rd Qu.: 60.93       3rd Qu.:300.7
##  Max.   :0.31026   Max.   :390.60       Max.   :302.2
##  reanalysis_avg_temp_k reanalysis_dew_point_temp_k
##  Min.   :296.2         Min.   :290.5
##  1st Qu.:298.6         1st Qu.:293.9
##  Median :299.5         Median :295.5
##  Mean   :299.6         Mean   :295.2
##  3rd Qu.:300.8         3rd Qu.:296.6
##  Max.   :302.2         Max.   :297.8
##  reanalysis_max_air_temp_k reanalysis_min_air_temp_k
##  Min.   :297.8             Min.   :293.3
##  1st Qu.:300.7             1st Qu.:296.5
##  Median :301.8             Median :297.6
##  Mean   :301.7             Mean   :297.5
##  3rd Qu.:302.9             3rd Qu.:298.7
##  Max.   :304.3             Max.   :299.9
##  reanalysis_precip_amt_kg_per_m2 reanalysis_relative_humidity_percent
##  Min.   :  0.00                  Min.   :70.64
##  1st Qu.:  8.57                  1st Qu.:75.24
##  Median : 19.10                  Median :77.50
```

```
##  Mean   : 26.75                  Mean   :77.54
##  3rd Qu.: 32.30                  3rd Qu.:79.64
##  Max.   :254.95                  Max.   :87.58
##  reanalysis_specific_humidity_g_per_kg reanalysis_tdtr_k
##  Min.   :12.36                         Min.   :1.657
##  1st Qu.:15.26                         1st Qu.:2.386
##  Median :16.90                         Median :2.671
##  Mean   :16.65                         Mean   :2.705
##  3rd Qu.:18.05                         3rd Qu.:3.000
##  Max.   :19.44                         Max.   :4.429
##  station_avg_temp_c station_diur_temp_rng_c station_max_temp_c
##  Min.   :23.31      Min.   :4.529          Min.   :26.70
##  1st Qu.:25.56      1st Qu.:6.029          1st Qu.:30.00
##  Median :27.04      Median :6.443          Median :31.70
##  Mean   :26.89      Mean   :6.499          Mean   :31.33
##  3rd Qu.:28.17      3rd Qu.:7.043          3rd Qu.:32.80
##  Max.   :30.03      Max.   :8.671          Max.   :35.00
##  station_min_temp_c station_precip_mm  total_cases
##  Min.   :17.80      Min.   :  0.00    Min.   :  1.00
##  1st Qu.:21.10      1st Qu.:  7.60    1st Qu.:  7.00
##  Median :22.80      Median : 20.90    Median : 13.00
##  Mean   :22.62      Mean   : 30.68    Mean   : 21.62
##  3rd Qu.:23.90      3rd Qu.: 41.00    3rd Qu.: 22.00
##  Max.   :25.60      Max.   :207.50    Max.   :170.00
```

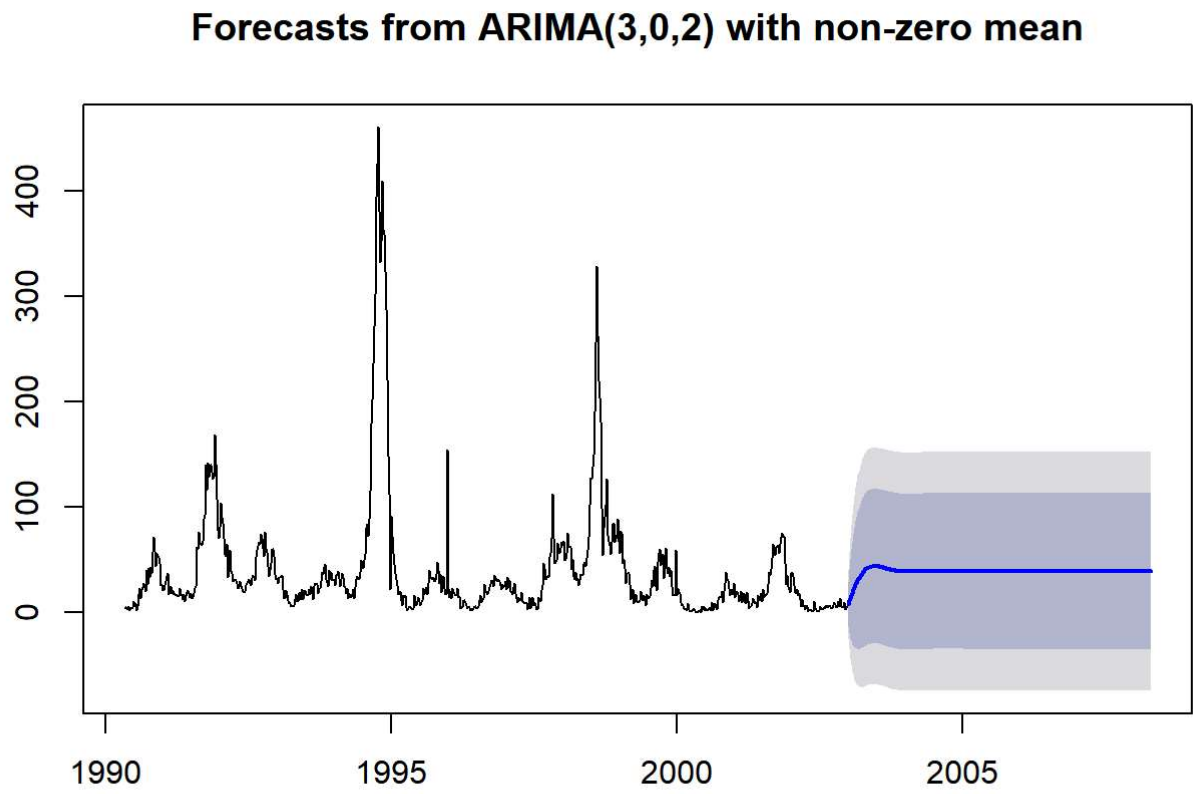
```
# Convert data to TS
sjtest.ts1 <- ts(sjdata.test.imputed,
                 freq=365.25/7,
                 start=decimal_date(ymd("2003-01-01")))
```

```
## Warning in data.matrix(data): NAs introduced by coercion
```

```
# Take lags of predictors
precipitation1 <- create.tslag(x = "precipitation_amt_mm"
                               ,sjtest.ts1)

dew.temp1 <- create.tslag(x = "reanalysis_dew_point_temp_k"
                          ,sjtest.ts1)

relative.humidity1 <- create.tslag(x = "reanalysis_relative_humidity_percent"
                                   ,sjtest.ts1)

avg.temp1 <- create.tslag(x = "station_avg_temp_c"
                          ,sjtest.ts1)

diur.temp1 <- create.tslag(x = "station_diur_temp_rng_c"
                           ,sjtest.ts1)

max.temp1 <- create.tslag(x = "station_max_temp_c"
                          ,sjtest.ts1)

min.temp1 <- create.tslag(x = "station_min_temp_c"
                          ,sjtest.ts1)
```

Evaluate the model accuracies

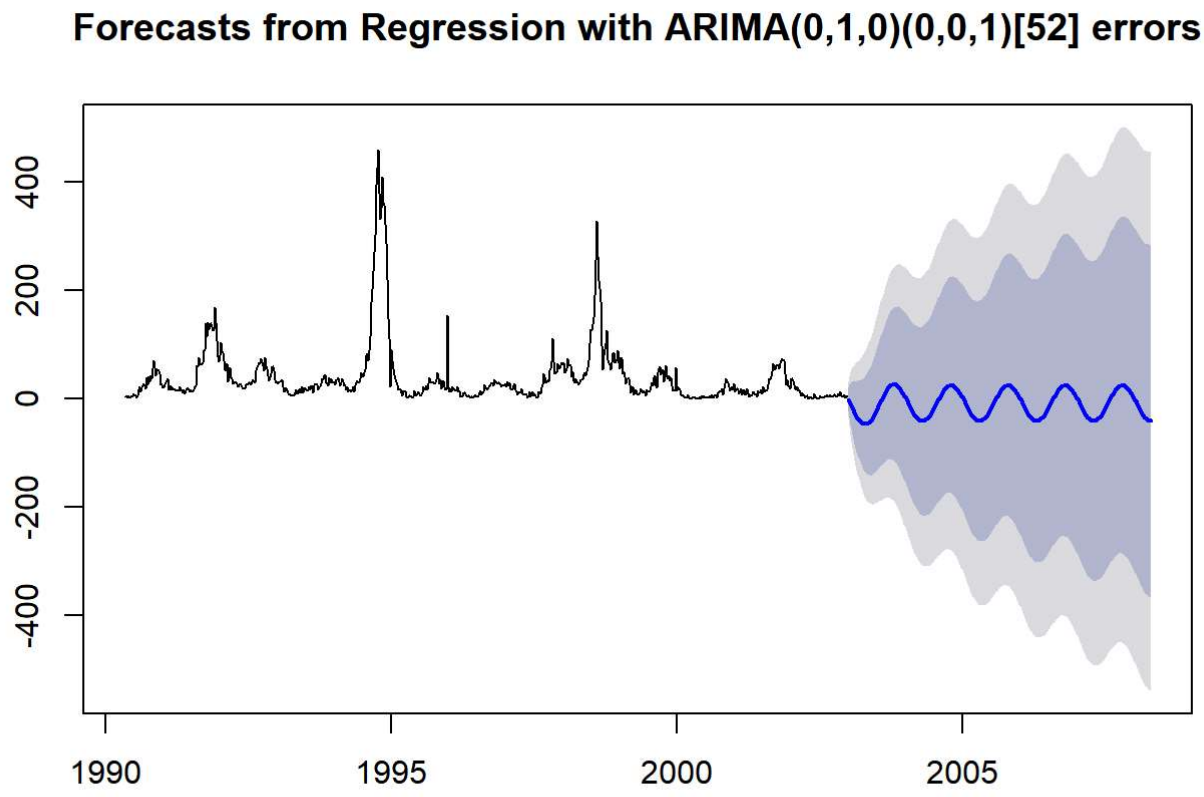
```
# Try ARIMA
fc0 <- forecast(sj.fit1,h=277)
plot(fc0)
```




```
accuracy(fc0$mean, as.integer(sjtest.ts1[,"total_cases"]))
```

```
##                 ME     RMSE      MAE       MPE     MAPE
## Test set -16.96754 31.63296 27.46105 -401.3438 412.3383
```

```
# Try SARIMA
fc1 <- forecast(fit1, xreg=fourier(sjtrain.ts, K=1, h=277))
plot(fc1)
```




```
accuracy(fc1$mean, as.integer(sjtest.ts1[,"total_cases"]))
```

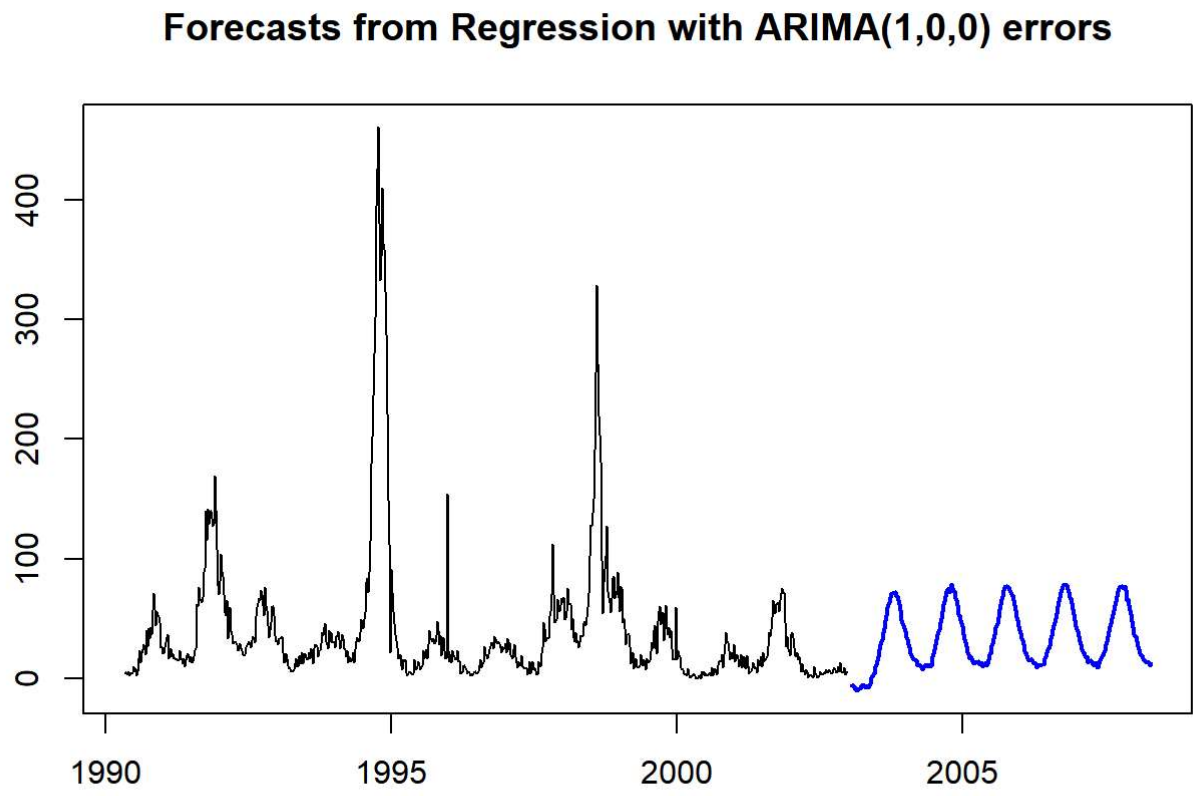
```
##                ME     RMSE    MAE      MPE     MAPE
## Test set 30.44953 39.88864 32.052 415.2451 431.4451
```

```
# Try SARIMAX with lag 4
fwf <- fourier(sjtrain.ts, K=2, h=277)
fc2 <- forecast(fit2, xreg=cbind(precipitation1[,4],relative.humidity1[,4]
                                 ,dew.temp1[,4],avg.temp1[,4]
                                 ,diur.temp1[,4],max.temp1[,4]
                                 ,min.temp1[,4],fwf), h =277)
```

```
## Warning in forecast.Arima(fit2, xreg = cbind(precipitation1[, 4],
## relative.humidity1[, : xreg contains different column names from the xreg
## used in training. Please check that the regressors are in the same order.
```

```
## Warning in forecast.Arima(fit2, xreg = cbind(precipitation1[, 4],
## relative.humidity1[, : Upper prediction intervals are not finite.
```

```
plot(fc2)
```



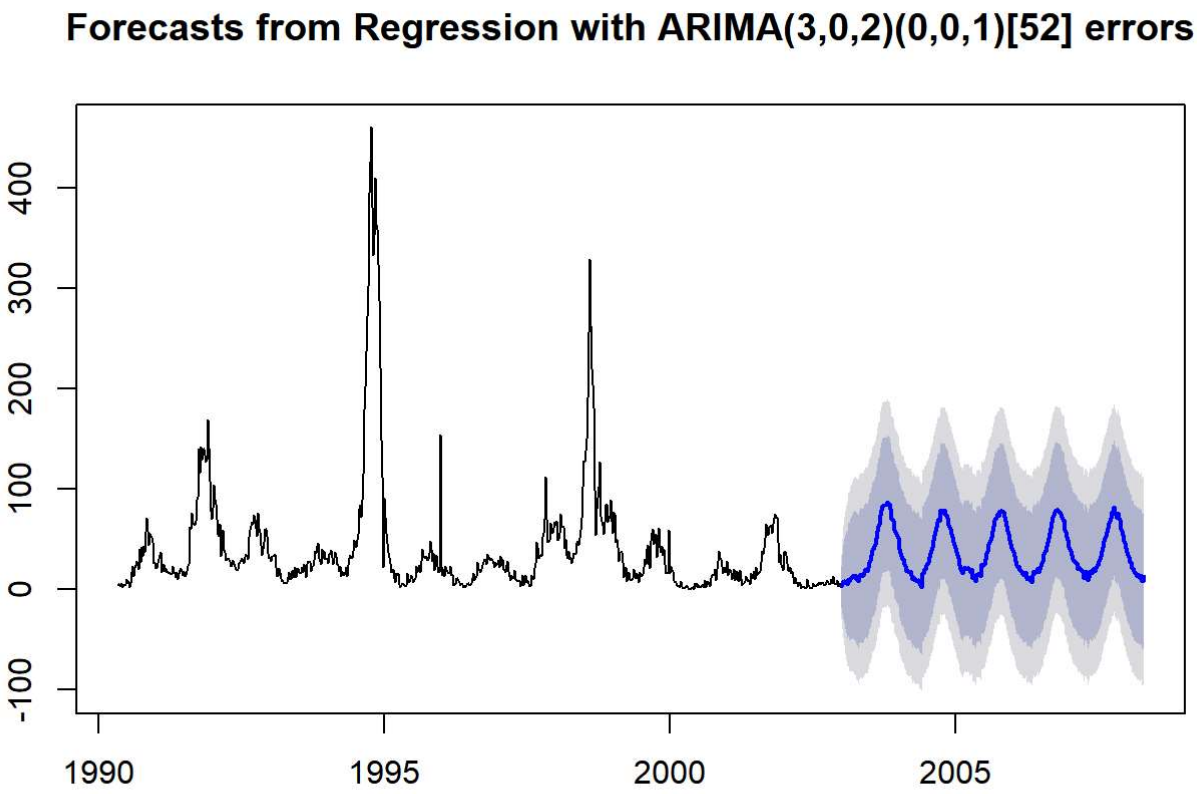

```
accuracy(fc2$mean, as.integer(sjtest.ts1[,"total_cases"]))
```

```
##                 ME     RMSE      MAE       MPE     MAPE
## Test set -12.63127 29.13512 22.56763 -173.1684 214.3513
```

```
# Try SARIMAX with lag 1
fc3 <- forecast(fit3, xreg=cbind(precipitation1[,1],relative.humidity1[,1]
                                 ,dew.temp1[,1],avg.temp1[,1]
                                 ,diur.temp1[,1],max.temp1[,1]
                                 ,min.temp1[,1],fwf), h =277)
```

```
## Warning in forecast.Arima(fit3, xreg = cbind(precipitation1[, 1],
## relative.humidity1[, : xreg contains different column names from the xreg
## used in training. Please check that the regressors are in the same order.
```

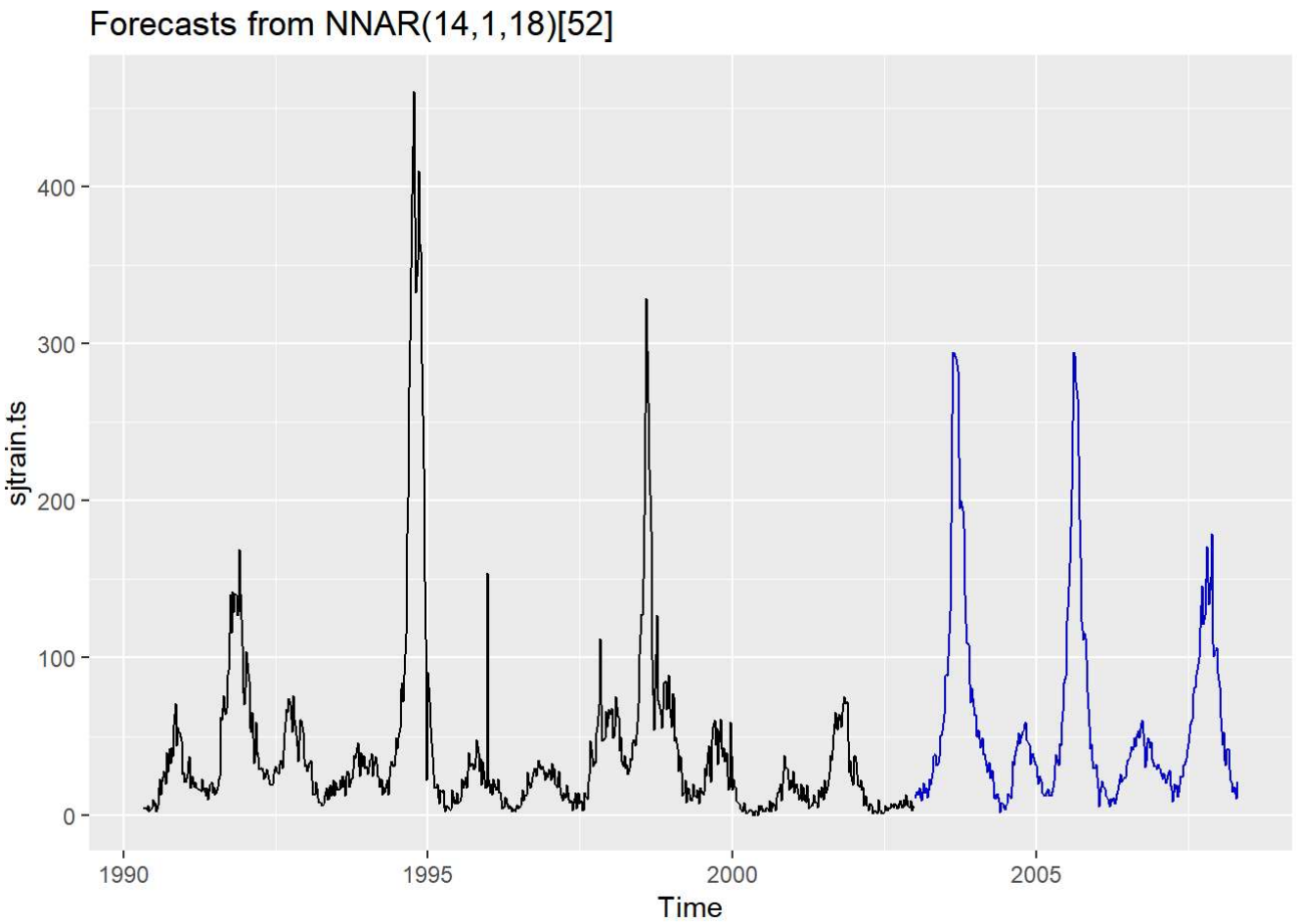
```
plot(fc3)
```




```
accuracy(fc3$mean,as.integer(sjtest.ts1[,"total_cases"]))
```

```
##                 ME     RMSE     MAE       MPE     MAPE
## Test set -15.71213 29.97731 23.2112 -203.4229 214.0362
```

```
# Try NN
fc4 <- forecast(fit4, xreg=sjtest.ts1[,5:23], h=277)
autoplot(fc4)
```




```
accuracy(fc4$mean,as.integer(sjtest.ts1[,"total_cases"]))
```

```
##                 ME     RMSE      MAE       MPE     MAPE
## Test set -38.47247 62.76566 39.41993 -293.5532 297.3995
```