Time series decomposition of air quality in Saigon, Vietnam

markdroes@gmail.com

February 13, 2018

*data was obtained through the US EPA (https://airnow.gov/index.cfm?action=airnow.global_summary#Vietnam$Ho_Chi_Minh_City)

```
#Reading in the 2017 hourly data
data.2016<-read.csv(file2016, sep = ';')[,c(3,8,9,11)]
data.2017<-read.csv(file2017)[,c(3,9,10,11)]
data.1.2018<-read.csv(file2018)[,c(3,9,10,11)]
data.2.2018<-read.csv(file.2.2018)[,c(3,9,10,11)]
#the 2016 data set is missing 743 obs. from 12/2016
dec.2016<-matrix(rep(c("2016-01",-999,-999,-999),743),nrow = 743, ncol = 4, byrow = T,
                 dimnames = list(c() , c("Date..LT.","AQI","AQI.Category","Raw.Conc.")))
#Bind data together row wise
data<-rbind(data.2016,dec.2016,data.2017,data.1.2018,data.2.2018)

#Data diagnostics on data$AQI
sum(data$AQI <0)
```

```
## [1] 1213
```

```
#1213 points with negative AQI (value actually -999), clearly errors.
#only 7% missing data, so we're doing well here
#Will impute values using 'imputeTS' pkg

#First step, set -999 to NuLL so the values are missing:
data$AQI[data$AQI == -999] <- NA
data$Raw.Conc.[data$Raw.Conc. == -999] <- NA

#Next, turn the data into a TS object to make it useable with 'imputeTS'
#Frequency indicates how many data points per cycle: here 365*24=8760
#Start = is when and where in the cycle we started. Here 2/11 2016 at 2pm, which is
  #day 42/365, and point 15 of that day. This becomes 2016+(41*24+15)/(24*365) = 2016.114
tsdata<-ts(data[,c(2,4)], frequency = 8760, start = 2016.114)
#Finally, the imputation (takes a while)
#imputed<-na.kalman(tsdata, model = "auto.arima")


#Write this new imputed file to a .csv (only need to do once)
#write.csv(imputed, file = "sgn_air_imputed.CSV")
```

First, let's a basic breakdown of the data. How many hours (data points) occurred where the air quality was "unhealthy for sensitive groups" or higher?

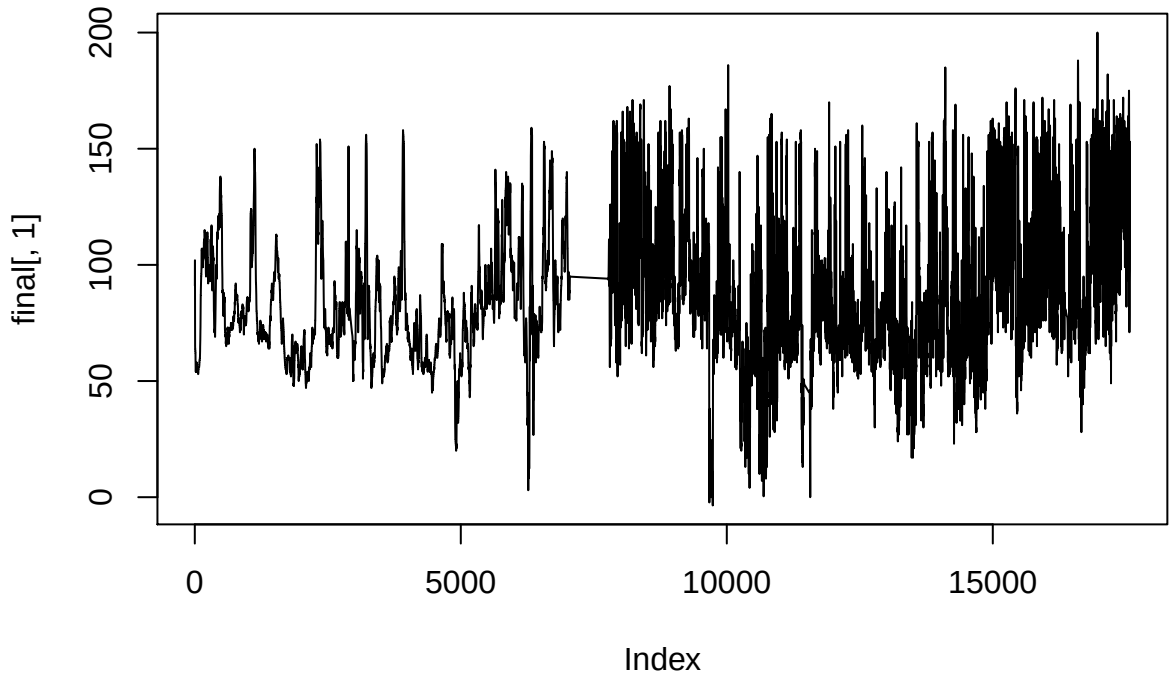
```
final<-read.csv("C:/Users/Mark/Documents/CSV/sgn_air_imputed.CSV")[,c(2,3)]
#This line counts the number of instances where AQI > 100
length(which(final[,1] > 100))
```

```
## [1] 4313
```

4313 hours over 2 years, which breaks down to ≈ 90 days a year.

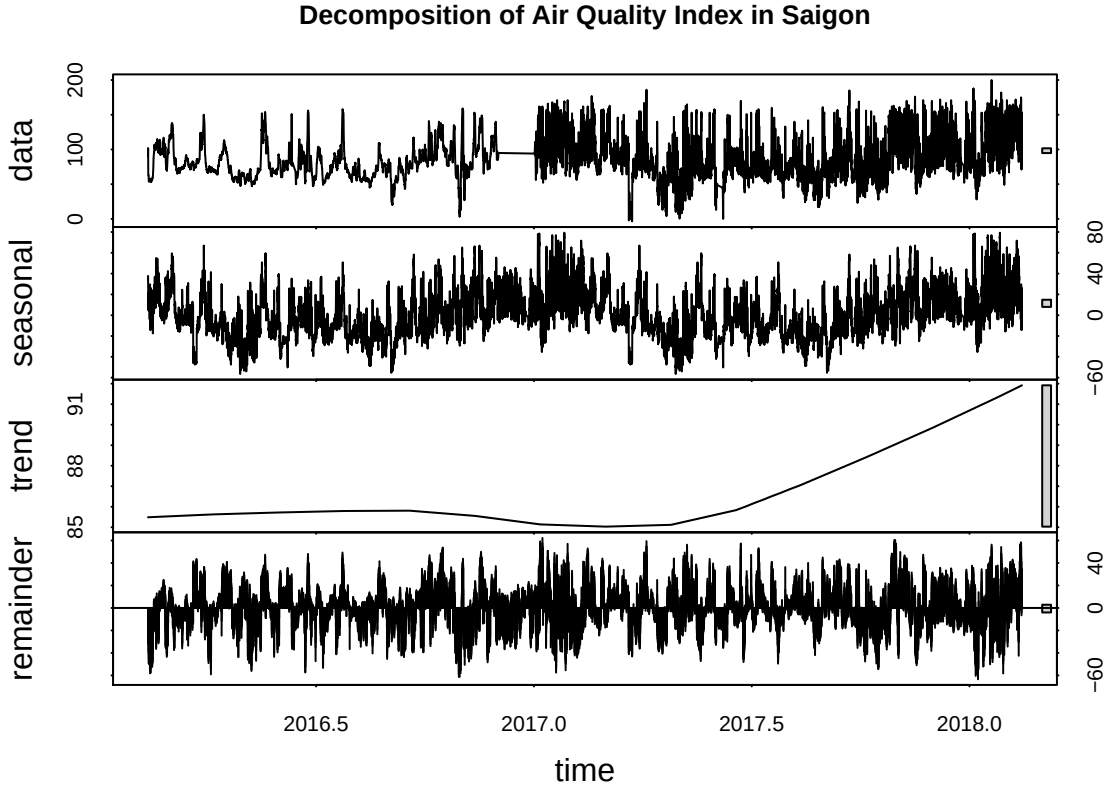
To see if there's any clear trends, let's next plot the AQI across the 2017 year.

```
plot(final[,1], type = 'l')
```



Unsuprisingly, we cannot infer much from this plot alone. The straight line part at about index 7500 was due to the necessity to impute the month of December, 2016. Next, I will perform a decomposition of this time-series data.

```
#stl() requires a time series class data set
imputed.ts<-ts(final,frequency = 8760, start = 2016.114)
#stl() will output a bunch of points, so we name it to a variable
decomp.aqi<-stl(imputed.ts[,1], s.window = "periodic")
#plots seasonality, trend, and remainder
plot(decomp.aqi, main = "Decomposition of Air Quality Index in Saigon")
```



Quite the trend uptick that begins about 4 months into 2017, right? Yes the model shows an increasing trend, but look closer at the units on the 4 graphs. Seasonal goes from -60 to 80 AQI, a range of 140. Remainder has a range of 100. Trend only has a range of 6! While this model shows an increasing AQI, and therefore an increase of pollution (which we would expect) – it is not as drastic as the figure shows. The next step to pursue is model diagnostics to ensure that our data is being properly represented.