Air Quality in Madrid

Prepared By: Mairuo (Michael) Liu, Hairun (Helen) Wang

Agenda

- Introduction & Business Problem
- Assumptions
- Data Properties & Transformation
- Time Series Model
- Model Selection
- Results
- Future Work

Poor air quality can be hazardous

- Air pollution is a tremendous problem in big cities, where health issues and traffic restrictions are continuously increasing
- There are some pollutants that cause immense disturbance to environment
- We will be exploring the **Air Quality in Madrid** and will forecast it with time series algorithms.

Take action on air pollution to save lives, and the planet, urges UN chief

Is Madrid about to reverse the traffic restrictions that solved its pollution problem?

Beijing's battle to clean up its air

LOS ANGELES RANKED HIGHEST IN U.S. FOR DEATHS LINKED TO AIR POLLUTION

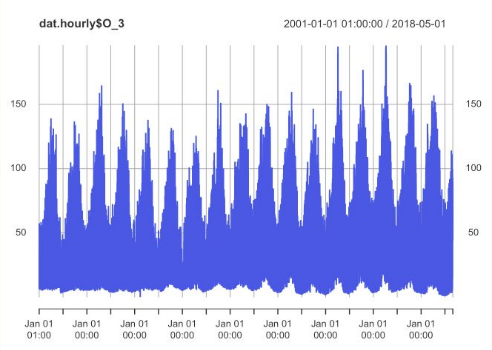
We have several assumptions before performing any analysis

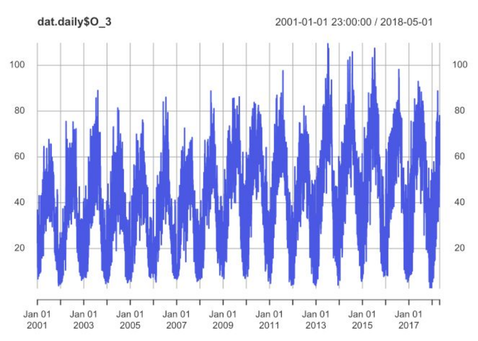
- Air quality data may have some seasonality caused by human activities
  - Special events, weekdays, economic/industrial life cycle
- Government policies or interventions can lead to diminishing trend in time series
- Time series of air quality data is autocorrelated

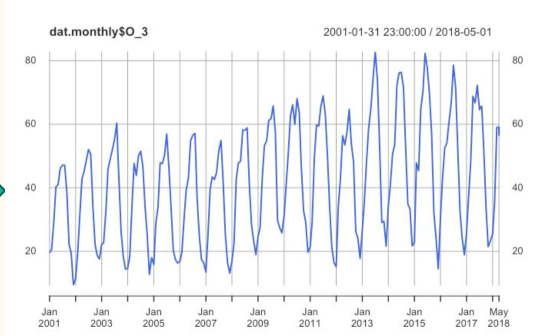
Data transformation is necessary due to properties of data set

- Multiple pollutants data observed from 2001/01 to 2018/04; ~4MM
- Only focused on 2 pollutants that cause threat to human health (also those 2 columns have less missing values):
  - **O3: Emitted by cars, power plants, industrial boilers, refineries, chemical plants**
  - **PM10: Formed from construction sites, unpaved roads, fields, smokestacks or fires**
- Each observation is based on hour
- 30 different observations measured at 30 different stations at every hour
- Impute missing values with linear interpolation

Aggregate data to reduce noises



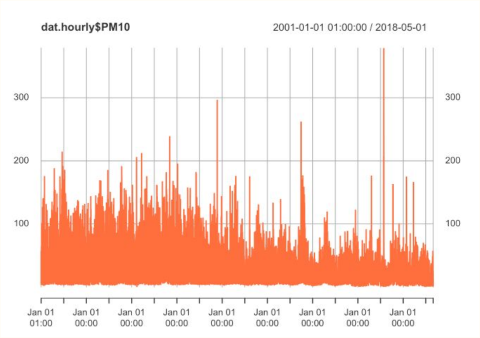


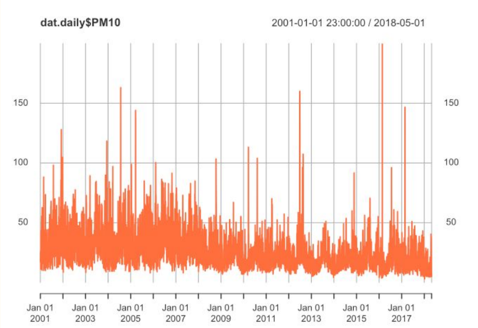


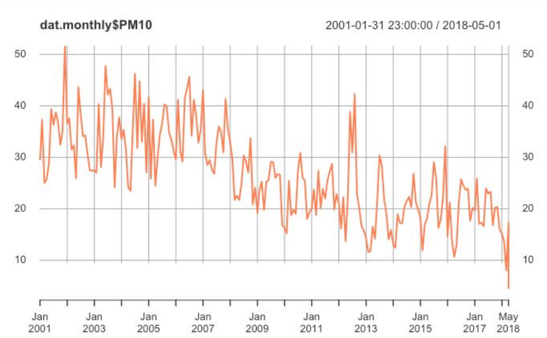
avg(08:00 at station1, station2, station3…)

max(average hourly data)

avg(max(Jun01, Jun 02, Jun 03))



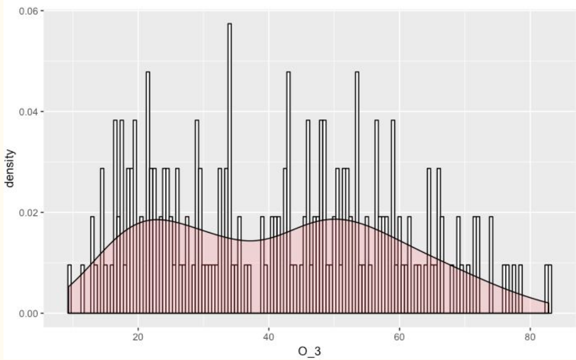


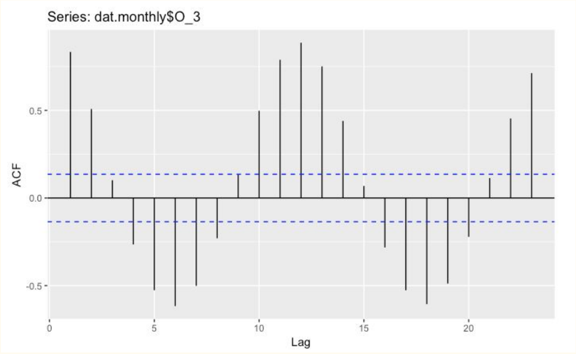


Data distribution is normal and no additional transformation needed

Ozone

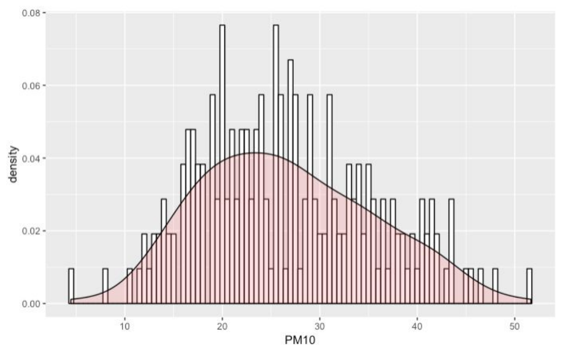
- Bimodal

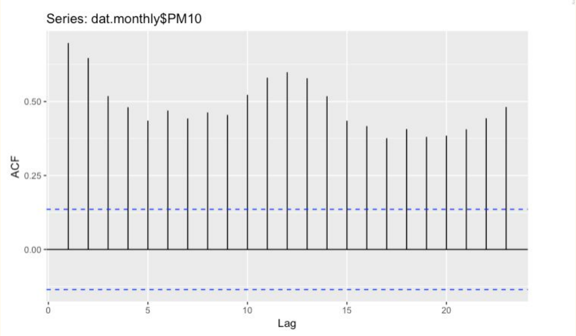




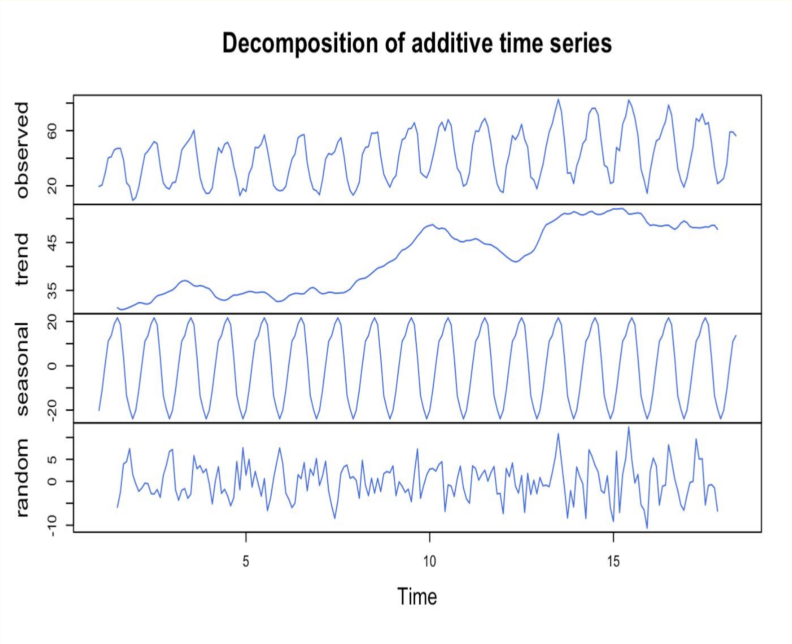
PM10

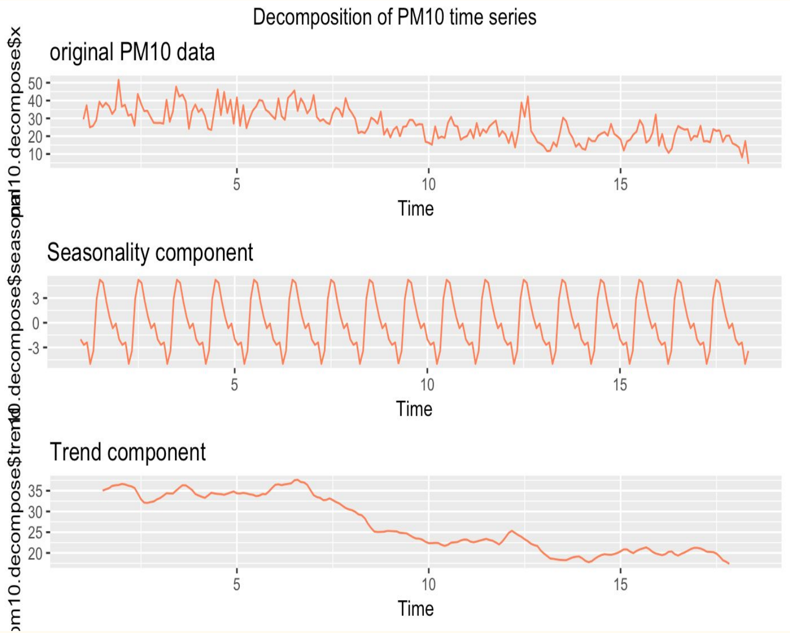
- Approx. Normal



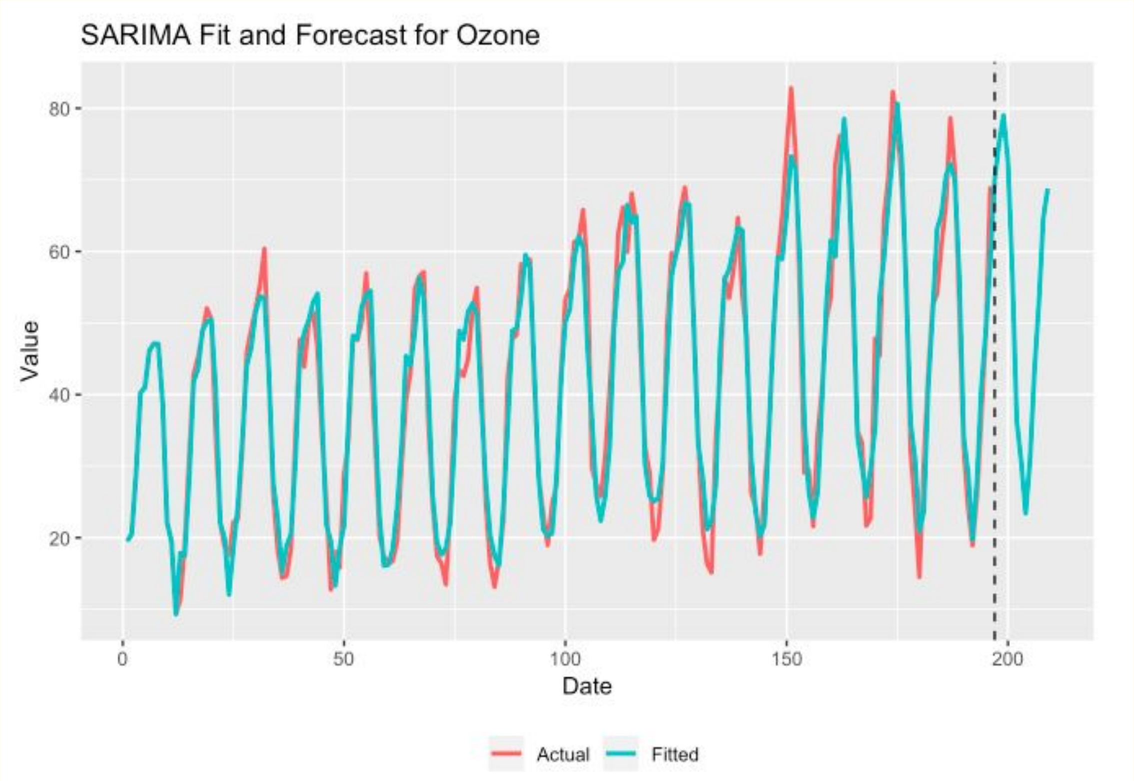


Decompose the time series to gain clearer insights





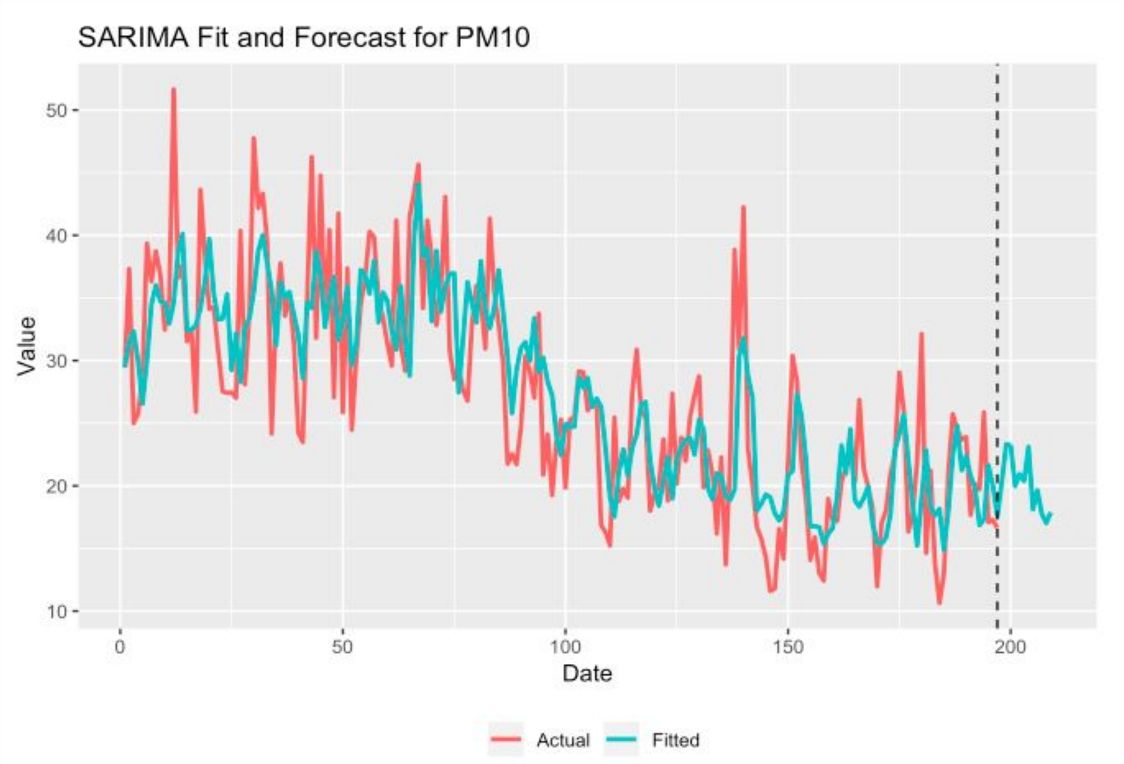
Seasonal ARIMA on Ozone Series



SARIMA(2,0,0)(0,1,1)[12] with drift

- ❖ Multiplicative Seasonality
- ❖ Stochastic trend

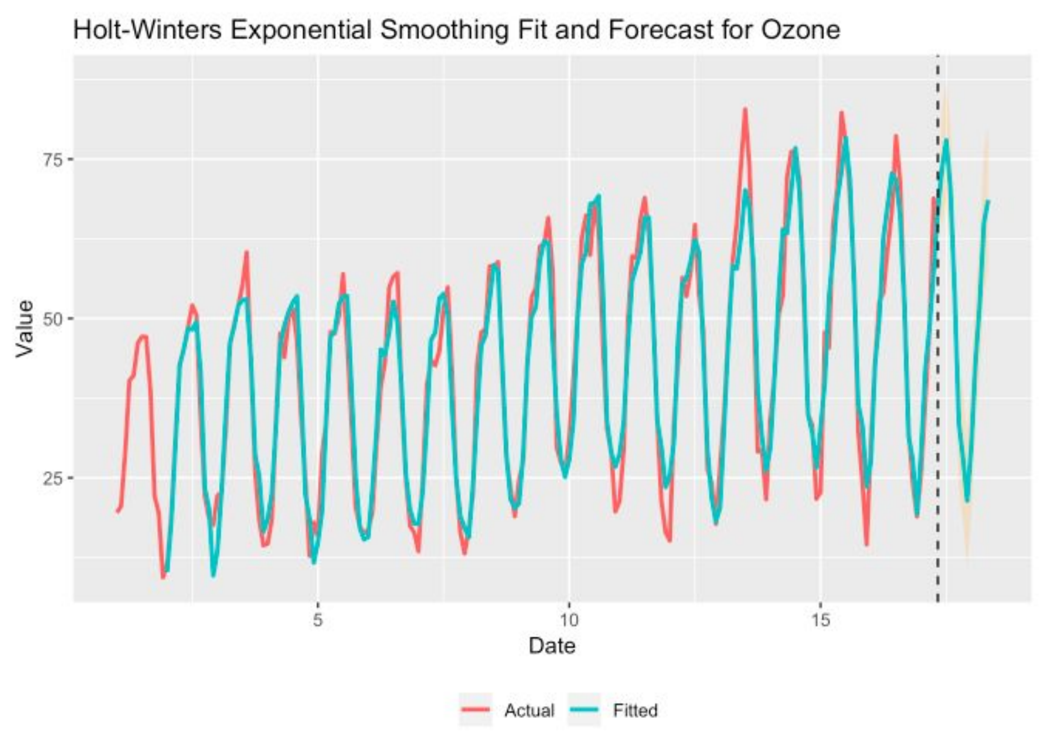
Seasonal ARIMA on PM10



SARIMA(4,1,1)(2,0,1)[12]

- ❖ Additive Seasonality
- ❖ Trend captured by I=1
- ❖ Performance not ideal (more on this eason)

Seasonal Holt-Winters and Exponential Smoothing State Space Model



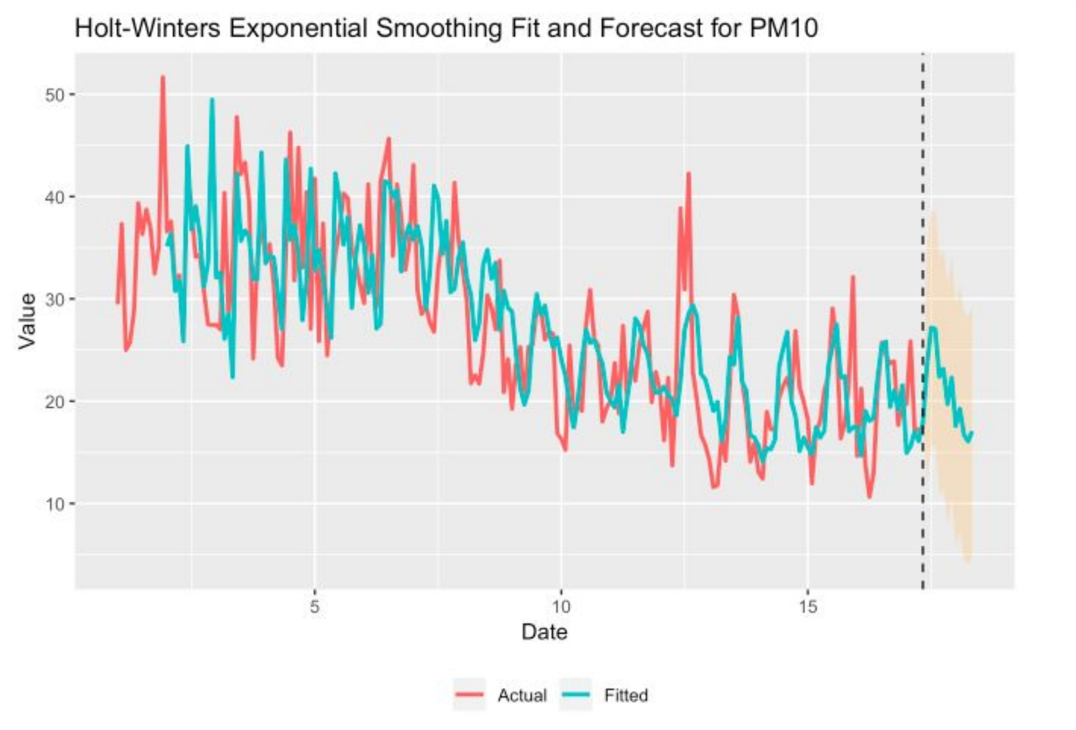
Exponential Smoothing State Space Model (called by ets()**)**

- simple exponential smoothing with additive errors, additive trend and additive season type (AAA)

Holt-Winters (called by HoltWinters()**)**

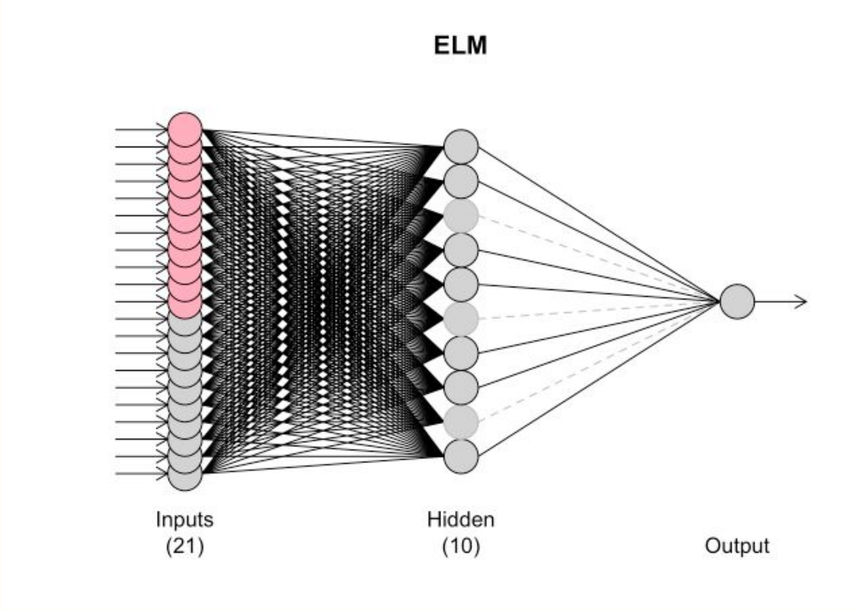
- Triple Exponential Smoothing
- Multiplicative seasonality
- Higher accuracy for Ozone

Seasonal Holt-Winters and Exponential Smoothing State Space Model



Does not perform as well as SARIMA.

Neural Network Model

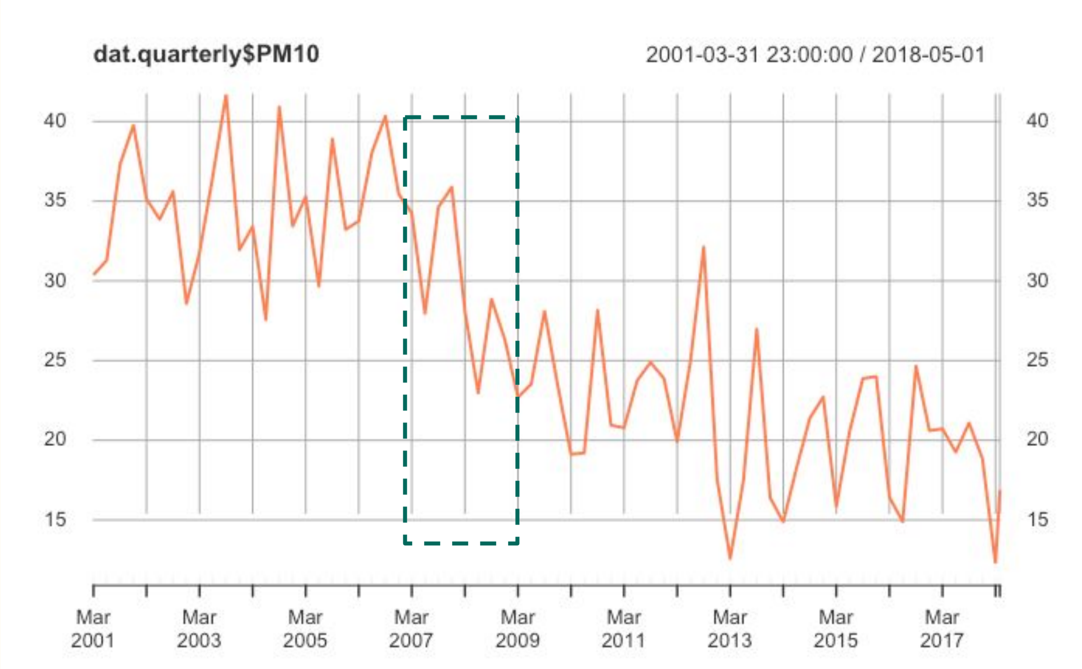


1 Hidden Layer

10 Hidden Nodes

Fully connected

A closer look at the quarterly plot of PM 10



- Obvious drop of PM10 value since 2008
- The plot can infer that this is a step response intervention with slope

Air quality and atmosphere protection. Law 34/2007, of 15 November 2007, on Air Quality and Atmosphere Protection (Law 34/2007) aims to prevent air pollution and to monitor and protect air quality. Law 34/2007 provides that both state and regional authorities must:

- Establish limits to air emissions.
- Establish air pollution prevention plans.
- Create an authorisation/notification system for potentially polluting activities that are not covered by the Recast Act on Integrated Pollution Prevention and Control (passed by Royal Decree Legislative 1/2016, of 16 December 2016) (IPPC Law).

*Source:https://uk.practicallaw.thomsonreuters.com/1-638-3346?transitionType=Default&contextData=(sc.Default)&firstPage=true&comp=pluk&bhcp=1

Hypothesis: PM10's drop happened because of intervention

```
pm.outliers <- 1*(seq(pm10.train)>=35)

pm.pulse <- arimax(pm10.train,order=c(4,1,1),seasonal=list(order=c(2,0,1),
                   period=12),xtransf=data.frame(pm.outliers),
                   transfer=list(c(0,0)), method='ML')

pm.pulse
```

```
NaNs produced
Call:
arimax(x = pm10.train, order = c(4, 1, 1), seasonal = list(order = c(2, 0, 1),
    period = 12), method = "ML", xtransf = data.frame(pm.outliers), transfer = list(c(0,
    0)))

Coefficients:
         ar1     ar2      ar3      ar4      ma1    sar1    sar2     sma1  pm.outliers-MA0
      0.2320  0.2188  -0.1245  -0.0508  -0.9176  0.5524  0.2076  -0.4589           1.5301
s.e.  0.0907  0.0780   0.0800   0.0797   0.0431  0.3339  0.1296   0.3484           3.1591

sigma^2 estimated as 27.56:  log likelihood = -605.47,  aic = 1228.95
```

Define the intervention time

Find level change and temp change as outliers

Feed data into ARIMAX model and plot the results

Hypothesis vs actual results

How do we validate the models?

Metrics:

- sMAPE
- AIC, BIC
- Box-Ljung Test for residuals

Method for CV:

- Hold-out set of 12 month worth of data

Here are the results model evaluation of each model:

O_3

	sMAPE	AIC	BIC
SARIMA	0.0689216	1097.53179765031	1113.6335767757
Exponential Smoothing State Space	0.0866228	1666.88809993114	1722.70256331968
Holt-Winter Exponential Smoothing	**0.0593831**	**1600ish**	**1700ish**
Neural Net	0.1118940	NA	NA

PM10

	sMAPE	AIC	BIC
SARIMA	**0.1357926**	**1229.18553402543**	**1258.68856595851**
Exponential Smoothing State Space	0.1494271	1707.09253011805	1756.34058604912
Holt-Winter Exponential Smoothing	0.1475189	1700ish	1700ish
Neural Net	0.1474199	NA	NA

Future Work

- Better implementation of Intervention Analysis
- Explore and perform more time series analysis for pollutants such as NO_2, SO_2, PM25 etc
- Combine weather, traffic data with air quality data and identify if any correlation or causality exists among those data sets -- regression model