BARC0130 : ADVANCED MATHEMATICAL MODELLING & ANALYSIS
MATLAB COURSEWORK : PROGRAMMING & DATA VISUALISATION

DATA PROCESSING

The data set was imported onto MATLAB and was found to have numerous missing values. In order for better visualisation and analysis of the data, these values were filled using the 'fillmissing' function on MATLAB with the 'makima' interpolation. 'Makima' was chosen as it's a cubic hermite interpolation method and would yield filled data with the least variance as it tajes more factors into consideration before estimating fill values. It must be noted that this involves the assumption that the data points are spatially or temporally related, which stands true as the data was collected at 15 minute intervals. However, since the values are approximated, it poses the limitation of these values being slightly different from the real data hence making the data less accurate.

Since the data measurements were taken every 15 mins for 92 days, it should be expected to have 8832 data points (96 each day). However, the data was observed to have 8836 (4 extra datapoints). A code was run to calculate the total data measurements of each day in order to find the day(s) with extra measurements. It was found that all days had an expected 96 measurements except 29th October, which had 100 measurements. The extra hour worth of measurements on this day could be associated with the daylight saving time.

In order to perform a more accurate analysis of the data, the 4 extra measurements need to be eliminated. This is done by taking a running mean of the 4 extra values with the 4 previous values.

$X_i = x_{i-1} + x_i + x_{i+1}/3$ where i = [1,3,5,6]

Where X_i corresponds to the new values and x_i correspond to old values. This was done by creating a 'datashrink' function. It must be noted that this merging process can be seen as a limitation as it is done by using the statistical tool 'mean' instead of the raw data which is not an extremely accurate representation of real data. It would have a certain deviation from the real values.

From the shrunk data, the NO_2 concentrations are plotted against time. This data is then separated into a weekly dataset by creating a 'data_x_weeks' function and the weekly concentrations of NO_2 are also plotted against time.

In order to visualise the data with eliminate the noise, the 'time series smoothening method' is applied which removes random variations within the data that may have been caused due to observational error. The exponential smoothening technique is chosen as it weighs the past observations using exponentially decreasing weights. This is done as follows:

$S_i = \alpha x_{i-1} + (1 - \alpha)\, S_{i-1}$ for $0 < \alpha \leq 1$ and $i \geq 3$ (S_1 = NaN, $S_2 = x_1$)

Where, S_i = exponentially weighted moving average

α = smoothening constant

i = observation index

x_i = dataset

There is no S_1. The speed at which the older observations lose their influence on the smoothed observation S_i is determined by the value of α. If α is close to 1, the damping effect is quick whereas, if α is close to 0, the damping effect is slow.

With the exponential smoothing technique, α is determined such that it will provide the optimal estimator in the form of the smoothed observation. This is done by creating a loop which takes all values of α = 0.1: 0.1: 0.9, and calculates the RMS of the estimation error and then chooses the value of α that minimises the RMS of the estimation error.

$$\text{RMS} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2} \qquad ; \varepsilon_i = S_i - \overline{X}$$

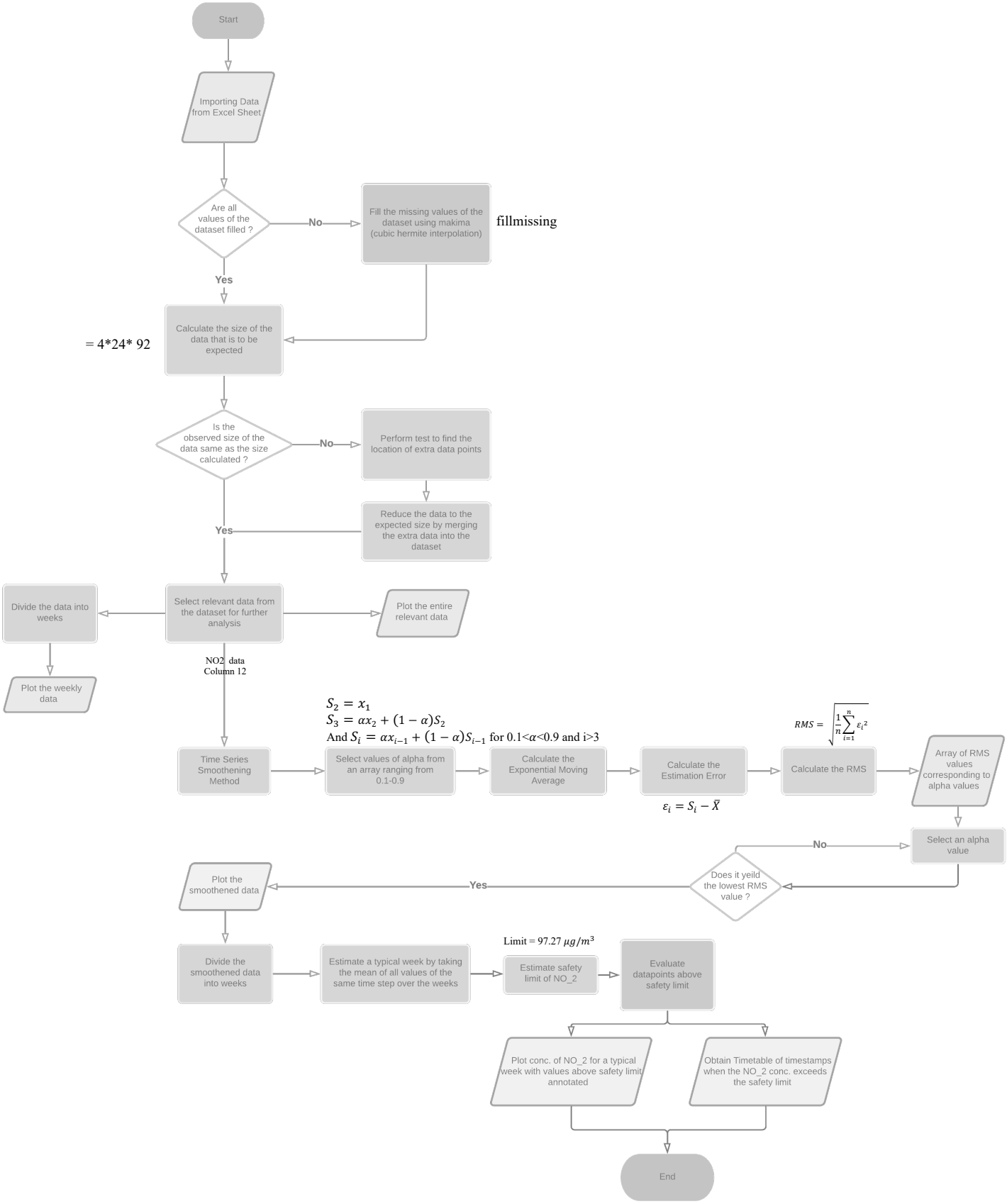
where n = no. of data points

$\overline{X}$ = mean of the data

It is inferred that for α = 0.1, the value of the RMS is the lowest, i.e., RMS = 26.8040 micrograms/m^3 . This is taken as the final α to obtain the smoothened data. This is done by creating a 'tsa_exponential' function. The smoothened data is plotted against time. This data is then separated into a weekly dataset by creating a 'data_x_weeks' function and the weekly smoothened are also plotted against time. It is observed that the 14th week has less data points than the other weeks and hence to obtain a typical week data from each time step over the weeks are averaged. The typical week obtained is then plotted.

A box plot is produced that helps analyse the distribution of data on each day. It should be noted that the plot shows the median but in this case where the data has already been smoothened out, the mean is a more reliable factor in determining the central tendencies of the data.

According to the air quality index, the safety limit of the amount of hourly NO_2 concentration is 0-50 ppb which corresponds to values above about 97.27 micrograms/m^3 being hazardous to health. A function 'safetylimit' is created to collect all values above the safety limit and create a timetable of their occurrence corresponding to the typical week plot.



DATA ANALYSIS

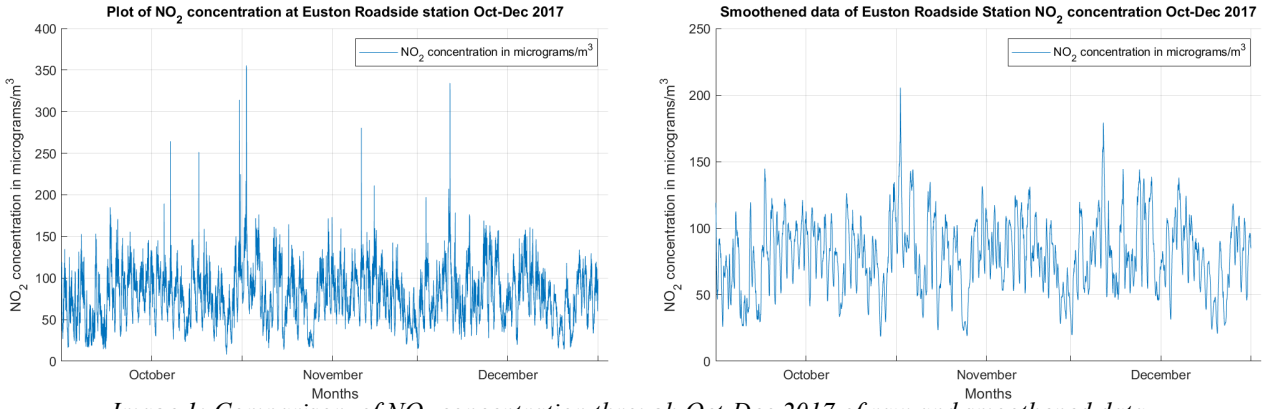


Image 1: Comparison of NO_2 concentration through Oct-Dec 2017 of raw and smoothened data

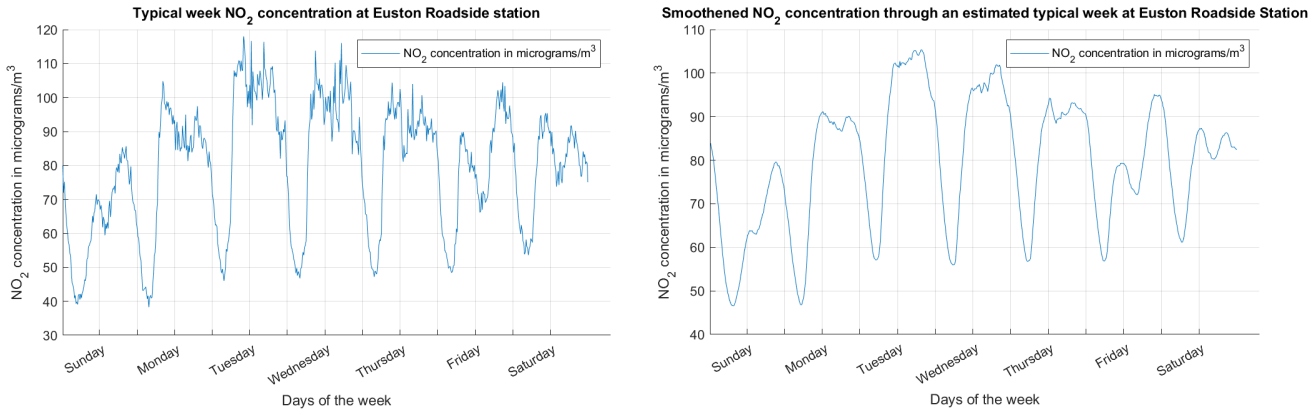


Image 2: Comparison of NO_2 concentration through a typical week of raw and smoothened data

As observed from *Image 1&2,* the graphs produced after the Time Series Smoothening Technique, produce a more clear trend to analyse the weekly patterns distributed through the data. It is observed that the smoothening successfully removed excessive noise from the data, that were potentially caused due to observational errors, and helped visualise a trend. This is evidence to prove that the purpose of smoothening was achieved.

Image 1 shows the daily fluctuation through Oct-Dec 2017. The highest spike in NO_2 conc. can be observed in the beginning of November. It can also be observed that the bulk of the values tend to fluctuate mainly between 50-100 micrograms/m^3. *Image 2* shows the data fluctuation through a typical week. It can be observed that the highest values are recorded on Tuesdays, followed by Wednesdays. The concentrations recorded on a Sunday are significantly lower than those on any other day of the week. It can be observed that the plot for each day assumes the shape of a bell curve implying that through the day the conc. steadily increases, reaches a peak (through the day) and then steadily decrease to a minimum (at dawn).

Image 3 is a box plot that shows the daily fluctuations of NO_2 concentrations through the estimated typical week. The plot marks the median of the data along with the 25th and 75th percentile data. This helps arrange the week in order of decreasing NO_2 conc. Tuesday >Wednesday > Thursday > Monday > Saturday > Friday > Sunday.

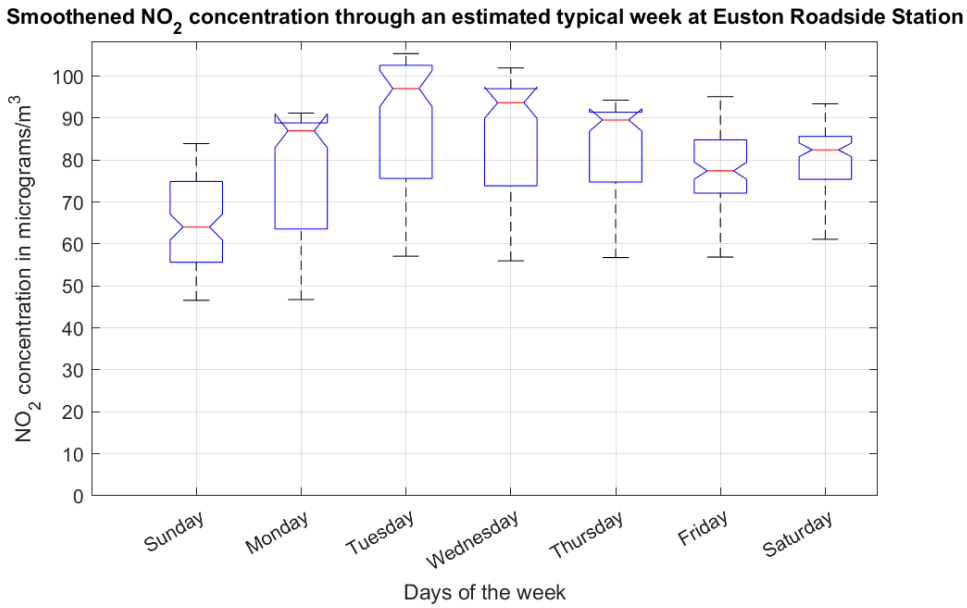


Image 3: Boxplot of smoothened data for typical week

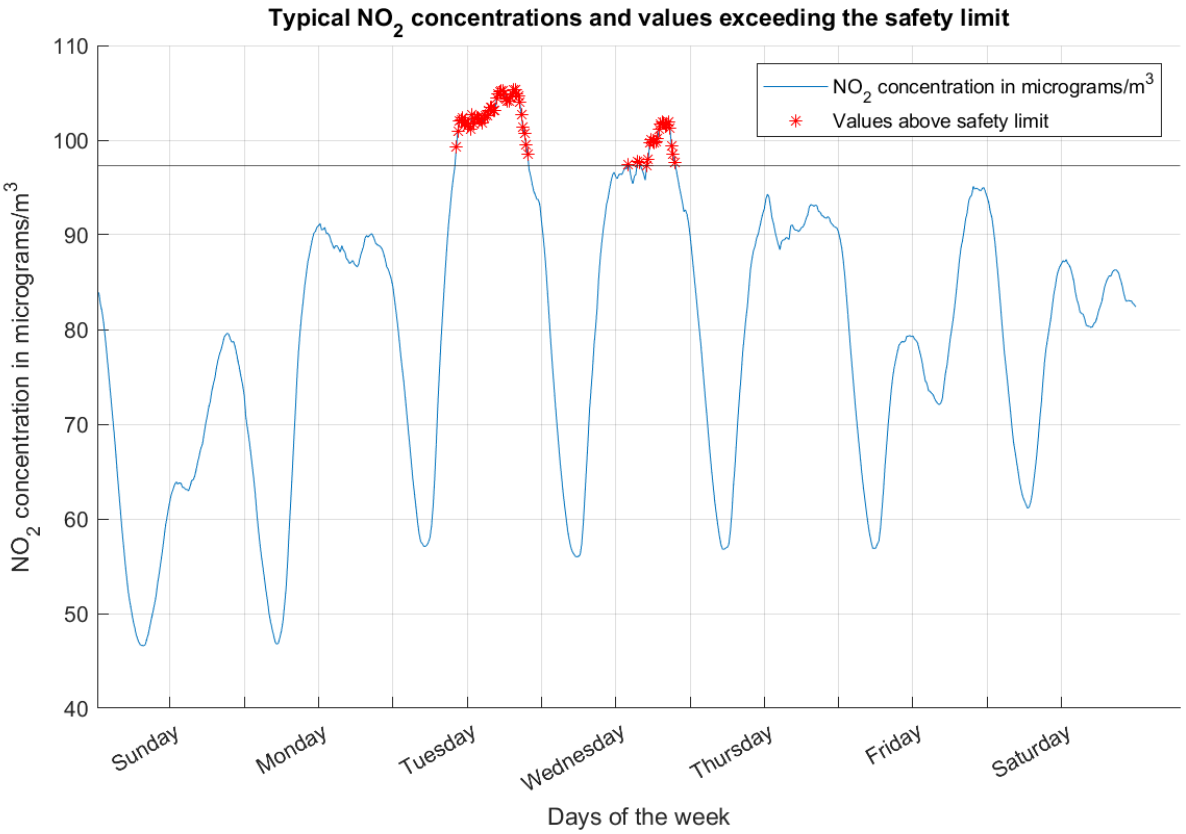


Image 4: NO_2 concentrations over a typical week showing values over safety limit

	1	2
1	Tuesday, 10:00 AM	Tuesday, 9:30 PM
2	Wednesday, 1:45 PM	Wednesday, 1:45 PM
3	Wednesday, 3:15 PM	Wednesday, 3:30 PM
4	Wednesday, 4:45 PM	Wednesday, 9:15 PM

Table 1: Timetable showing time intervals in the typical week with NO_2 conc. over safety limit

Figure 4 is a more comprehensive plot that shows the typical week of NO_2 concentration distributions while also visualising the data points wherein the concentration exceeds the safety limit of 92.27 micrograms/m^3. *Table 1* is a time table correlating with the graph above that shows the time intervals within which the NO_2 levels exceed the safety limit. Column '1' represents the beginning of the interval and Column '2' represents the end of the interval. It shows that the NO_2 levels on Wednesdays and Tuesdays are unhealthy through its day time duration. It is safer to go out during the mornings or late at night.

The overall concentration of NO_2 in the air for a typical week is higher on weekdays than on the weekends. Considering its location, it is very likely that the major emission is generated during working hours.

RESULTS AND CONCLUSION

The model helps deduce that it is safe to go out through the week, except it is advised to only step out during early mornings and after 9:30 at night during Tuesdays and Wednesdays.

It is however essential to note when considering the accuracy of the model that the period of measurement was very limited as the data was provided for just the quarter of one year. It can be deduced that there would be a seasonal change in these measurements considering the fact that temperature and other environmental parameters also effect the NO_2 concentration levels. Further, analysis of many years of data rather that just one would also increase the models accuracy. Lastly, the reliability of these observations cannot be ascertained and there is scope for observational error even though it was attempted to eliminate these using various statistical and mathematical tools.

REFERENCES

1. US Environmental Protection Agency. Air Quality Guide for Nitrogen Dioxide [Internet]. 2011 Feb. Available from: https://www.airnow.gov/sites/default/files/2018-06/no2.pdf
2. WKC group. Micrograms per Cubic Metre / Parts per Billion Converter [Internet]. WKC Group. [cited 2021 Dec 16]. Available from: https://www.wkcgroup.com/tools-room/micrograms-per-cubic-meter-parts-per-billion-converter/