By: Arun Chandra Sharvirala

**Analysis on Air Quality Dataset from UCI**

**About the Data:**

**Dataset Name**: Air Quality UCI

**Dataset Source**: http://archive.ics.uci.edu/ml/datasets/Air+Quality#

**Dataset Information:**

The dataset contains 9358 instances of hourly averaged responses from an array of 5 metal oxide chemical sensors embedded in an Air Quality Chemical Multisensor Device. The device was located on the field in a significantly polluted area, at road level,within an Italian city. Data were recorded from March 2004 to February 2005 (one year)representing the longest freely available recordings of on field deployed air quality chemical sensor devices responses. Ground Truth hourly averaged concentrations for CO, Non Metanic Hydrocarbons, Benzene, Total Nitrogen Oxides (NOx) and Nitrogen Dioxide (NO2) and were provided by a co-located reference certified analyzer. Evidences of cross-sensitivities as well as both concept and sensor drifts are present as described in De Vito et al., Sens. And Act. B, Vol. 129,2,2008 (citation required) eventually affecting sensors concentration estimation capabilities. Missing values are tagged with -200 value.

**Dataset Attribute Information:**

0 Date (DD/MM/YYYY)

1 Time (HH.MM.SS)

2 True hourly averaged concentration CO in mg/m^3 (reference analyzer)

3 PT08.S1 (tin oxide) hourly averaged sensor response (nominally CO targeted)

4 True hourly averaged overall Non Metanic HydroCarbons concentration in microg/m^3 (reference analyzer)

5 True hourly averaged Benzene concentration in microg/m^3 (reference analyzer)

6 PT08.S2 (titania) hourly averaged sensor response (nominally NMHC targeted)

7 True hourly averaged NOx concentration in ppb (reference analyzer)

8 PT08.S3 (tungsten oxide) hourly averaged sensor response (nominally NOx targeted)

9 True hourly averaged NO2 concentration in microg/m^3 (reference analyzer)

10 PT08.S4 (tungsten oxide) hourly averaged sensor response (nominally NO2 targeted)

11 PT08.S5 (indium oxide) hourly averaged sensor response (nominally O3 targeted)

12 Temperature in Â°C

13 Relative Humidity (%)

14 AH Absolute Humidity

**Analysis:**

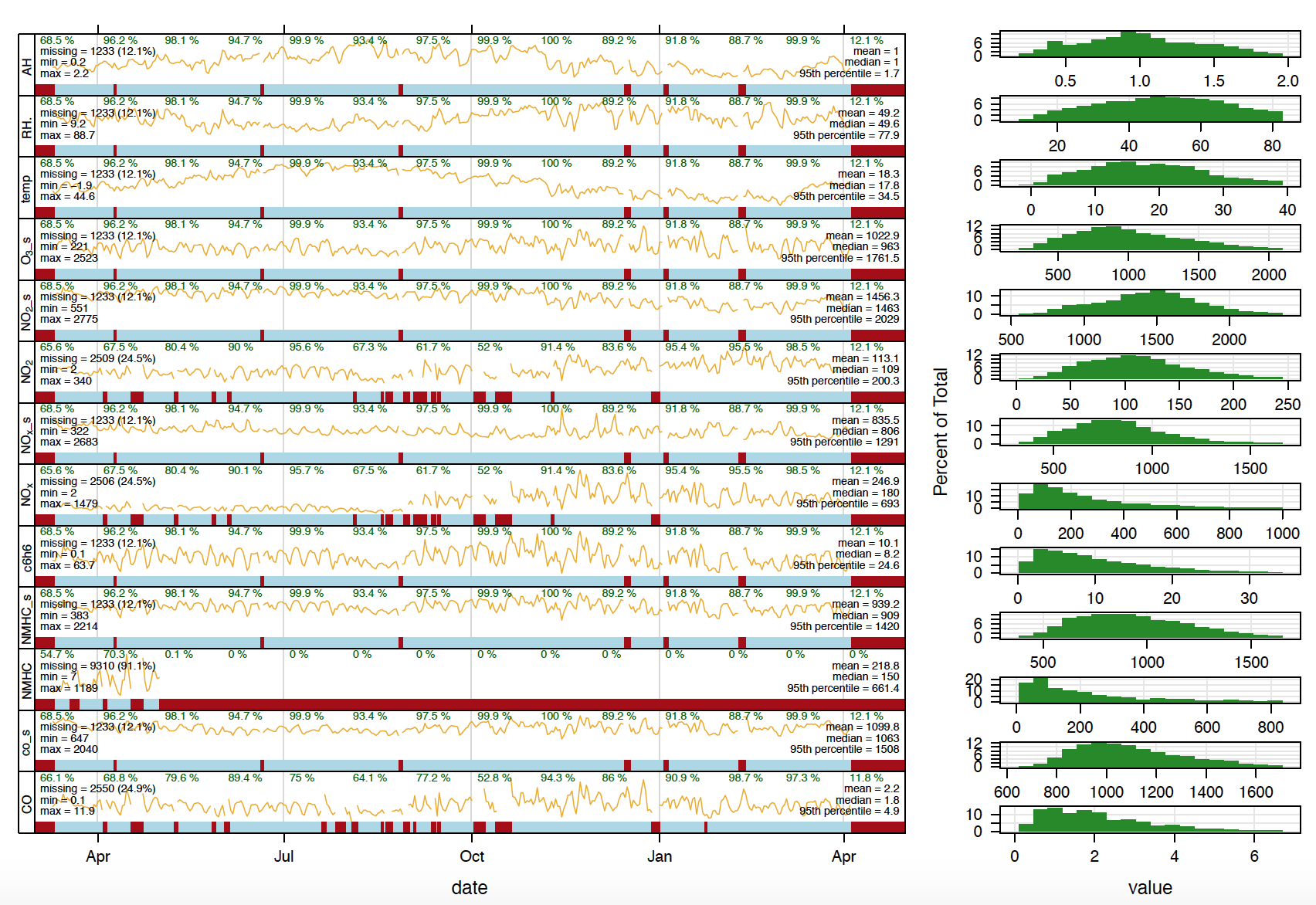
*Packages required:* openair, hydroGOF, MLmetrics, caTools, e1071, plotrix

Analysis of the city’s pollution levels with the ground true features.

The data collected by the sensor is denoted by ‘\_s’ after the targeted pollutant in the dataset for the above summary plot. Upon only considering the sensor data, there might be an anomaly in the sensor as there are missing values in all the data collected by the sensor exactly at the same intervals. Additionally, we can observe in the above plot that more than 90% of the ground true NMHC pollutant data is missing. We can also observe the distribution of each pollutant on the right side of the summary plot with histogram notation.

In the dataset, I have changed the feature names of the 5 metal oxide sensors to ‘\_s’ at the ending of their nominally targeted pollutant.

From the below summary plot, we can check for the percentage of missing values in each of the feature, mean and median of each feature. We can also check for the distribution of each feature in terms of histograms.

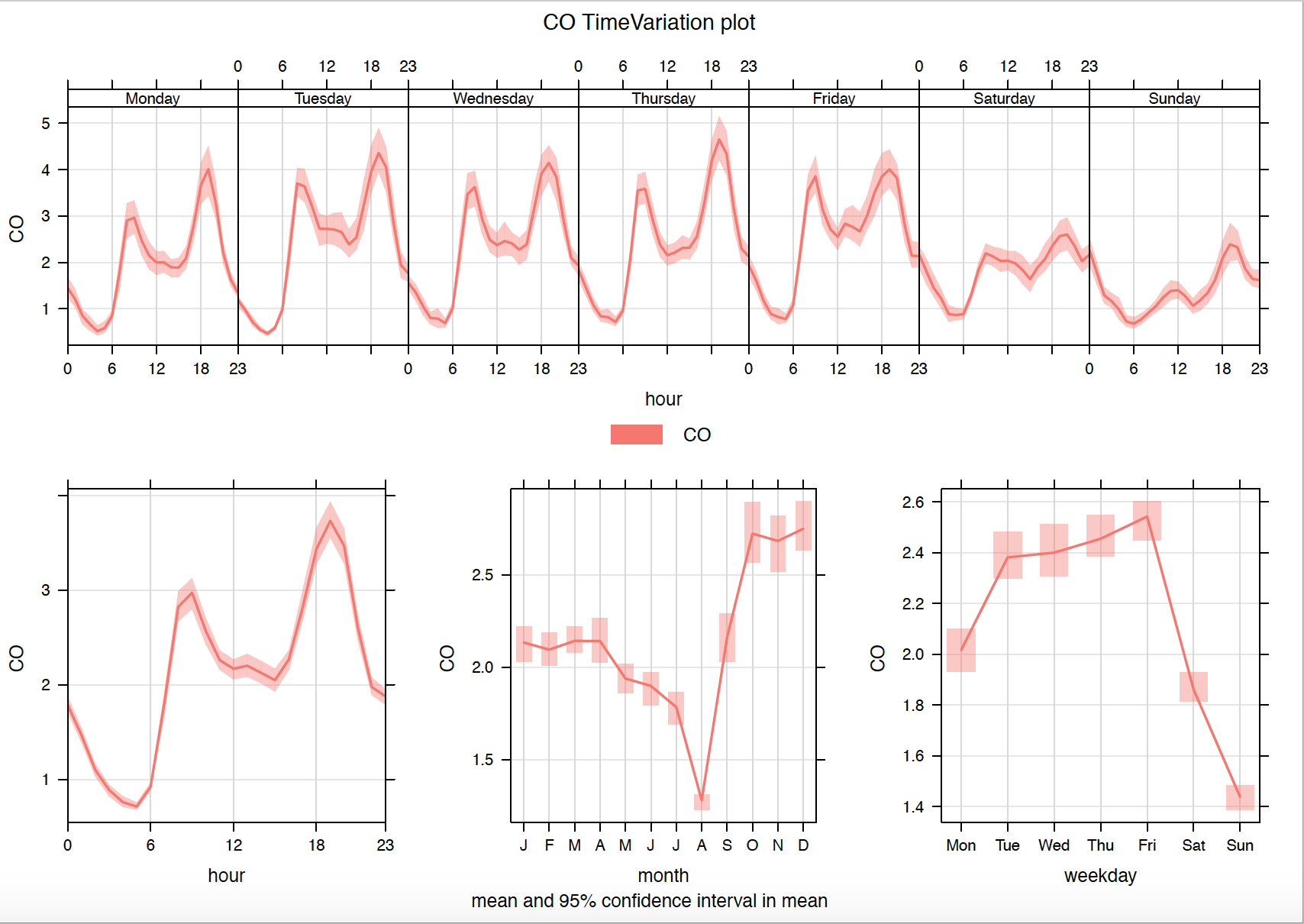


**TimeVariation Plot:**

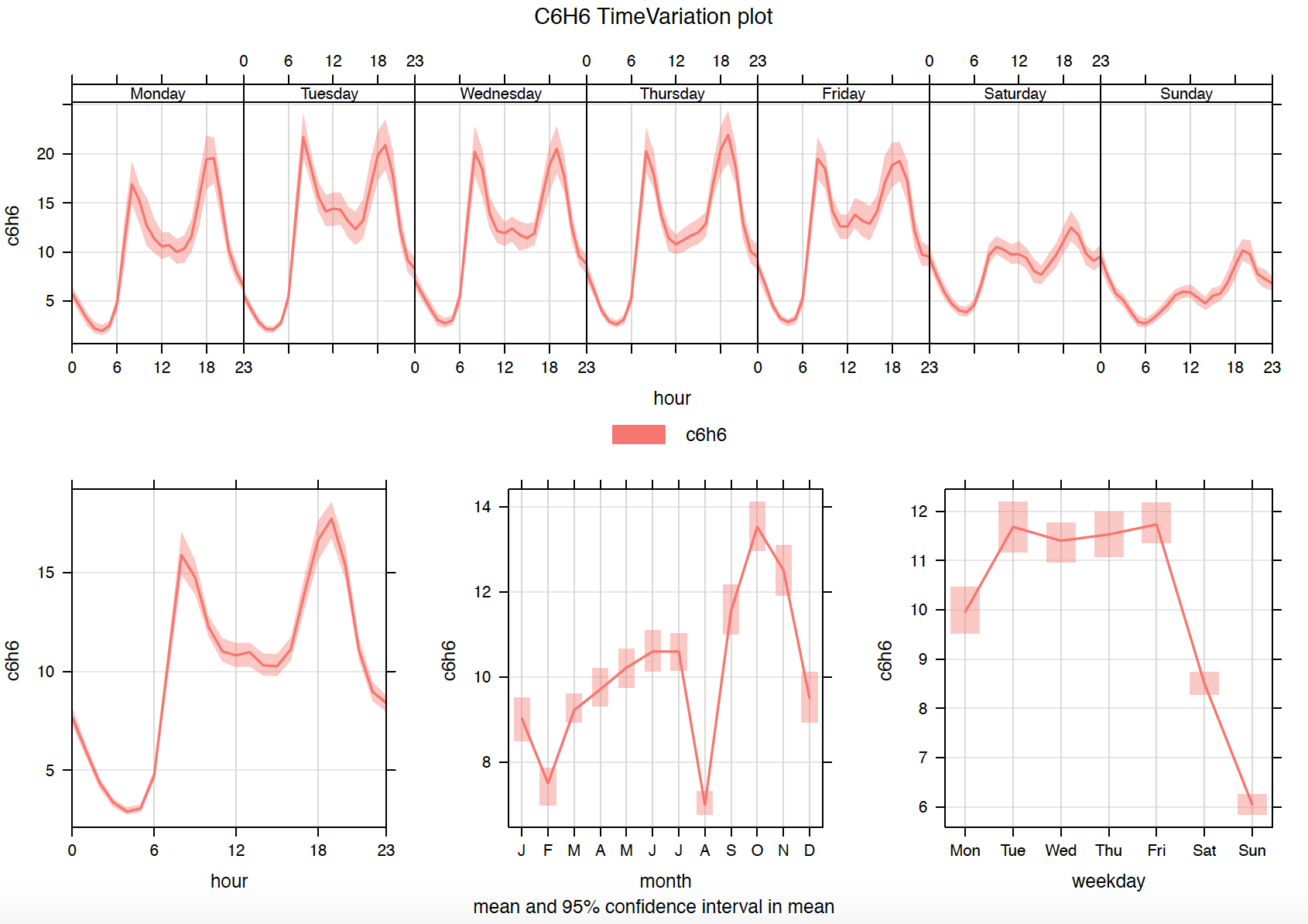
Time variation plot is plotted against every month, weekday and hour. As expected all the pollutant’s levels are rather high during the weekdays than on weekends and there are two sets of peak hours, one in the morning and one in the evening. Some pollutants concentration levels are same for morning and evening peak hour time frames, while other pollutant’s concentration levels are high in the evening than in the morning peak hour time frame.

From my intuition, I expected the pollutant levels would be higher in the summer season compared with any other season. But, from the below plots we can observe that the pollution levels are relatively high in winter season than any other season compared with. Astonishingly, august month has the lowest recorded pollutants concentration level.

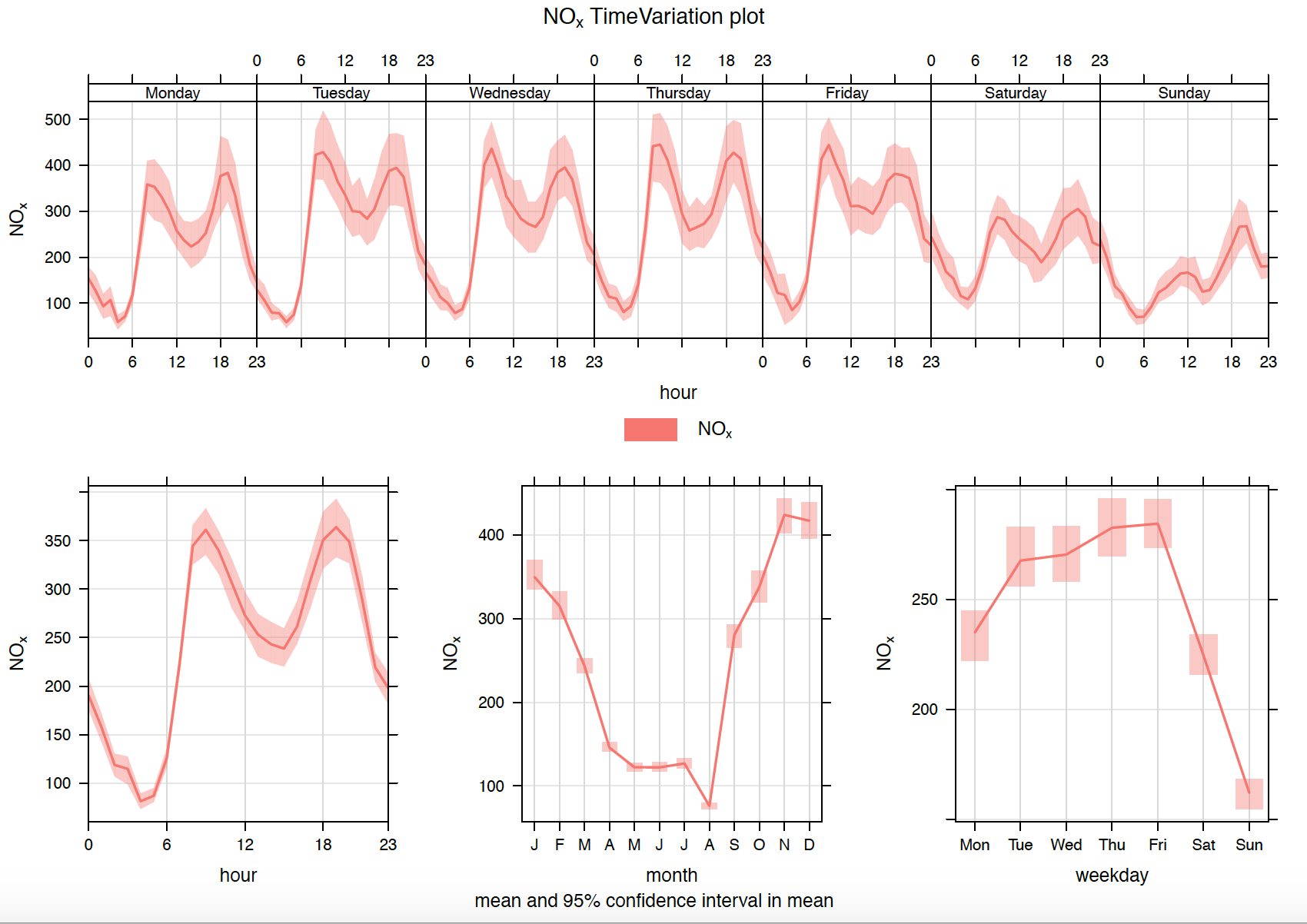
*Carbon Monoxide (CO):*



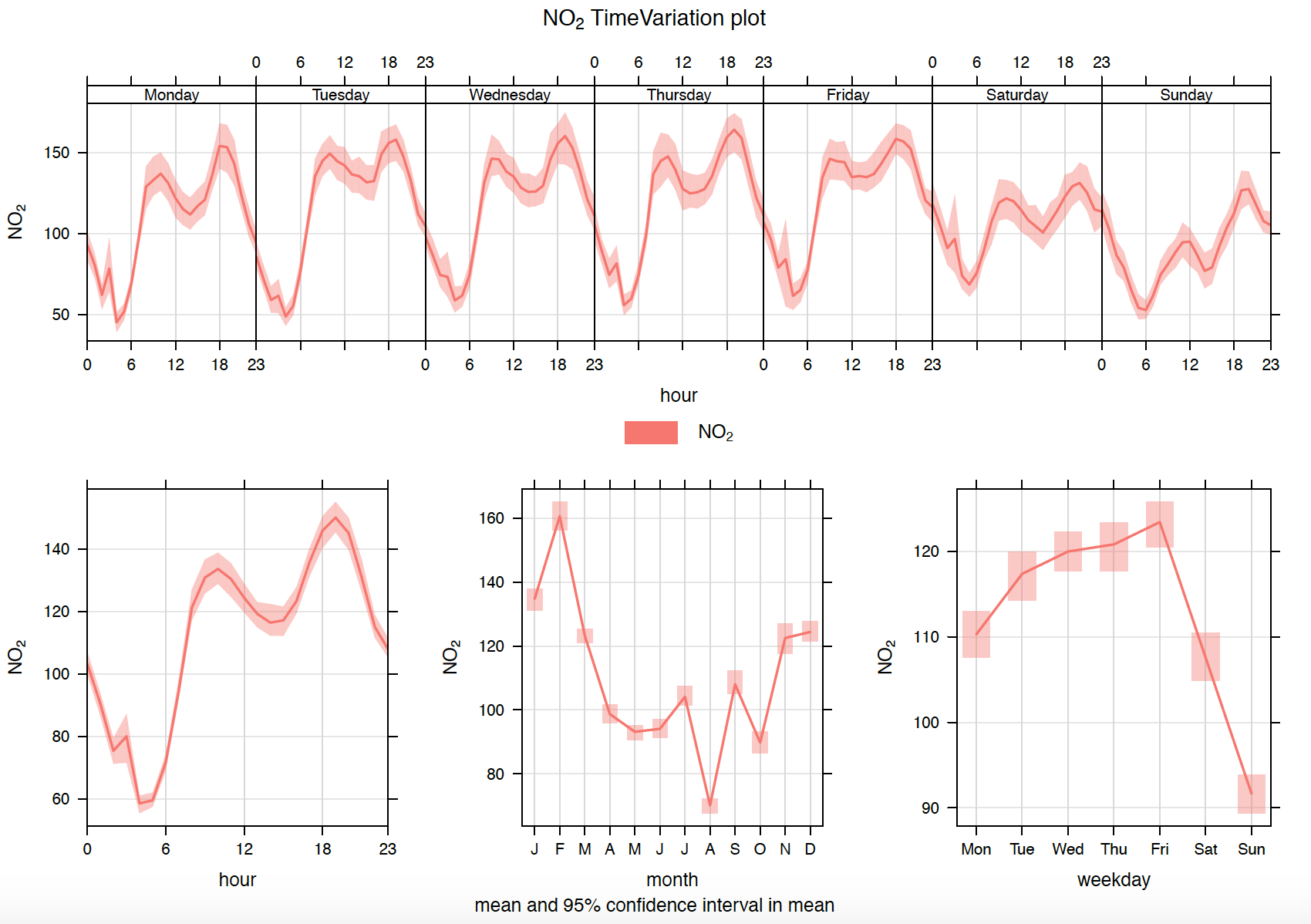
*Benzene (C6H6):*



*Nitrogen Oxides (NOx):*



*Nitrogen Dioxide (NO2):*

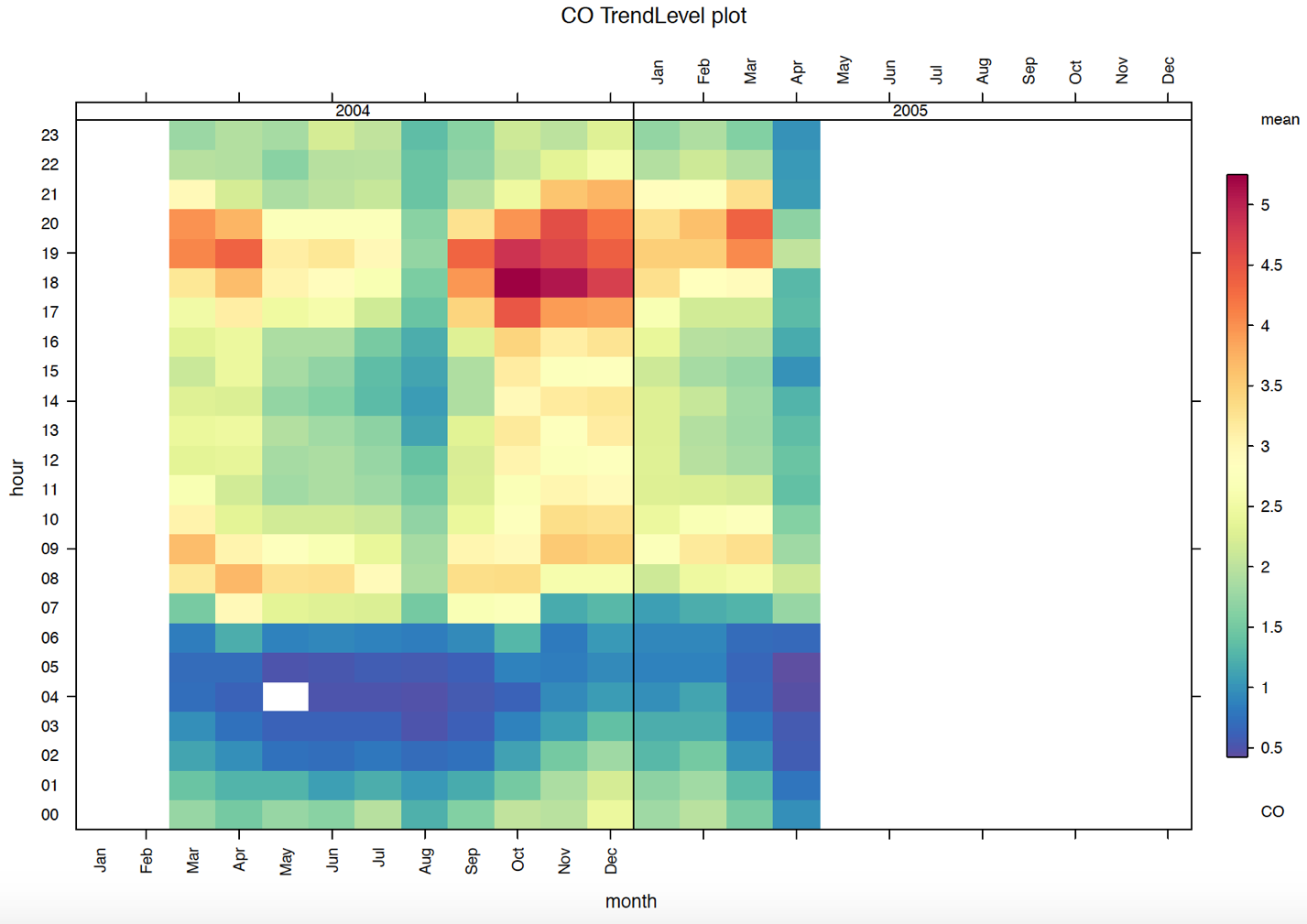


**TrendLevel Plot**:

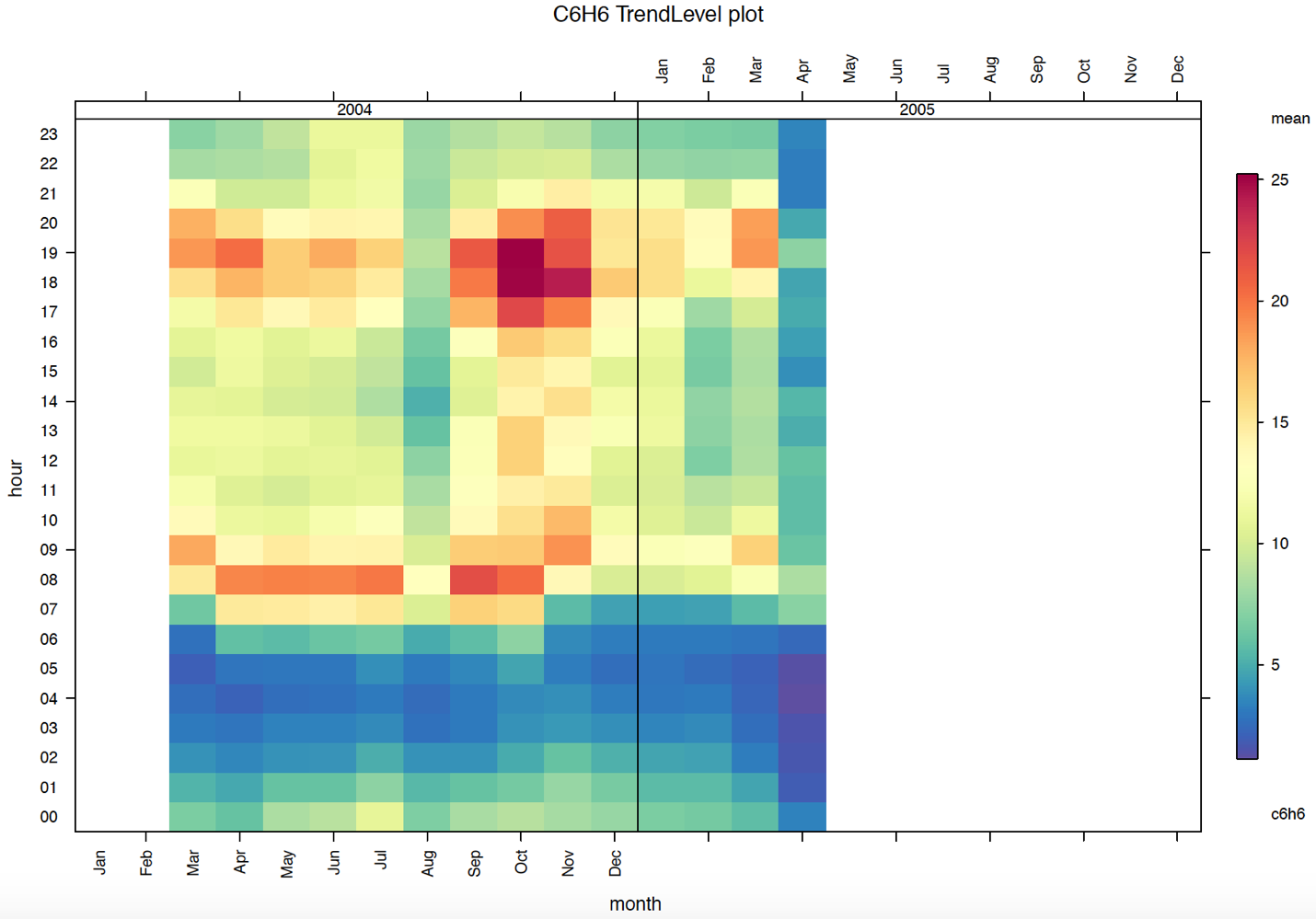
TrendLevel plot packs the whole data into a single plot as a heat map. X – axis represents the month, while Y – axis represents the hour. We can clearly observe the concentration of each pollutant with respect to each hour in a day. The concentration of all the pollutants is higher in the mornings (08:00 AM to 12:00 AM) and in the evenings (06:00 PM to 09:00 PM). In addition, the concentration of the pollutants is highest at 08:00 AM in the morning and 06:00 PM in the evening. Blank or white spaces in the heat maps is the missing data.

Nitrogen oxides have a higher concentration during the Autumn and winter seasons, while carbon monoxide and the benzene concentration is higher in the autumn and starting of the winter season.

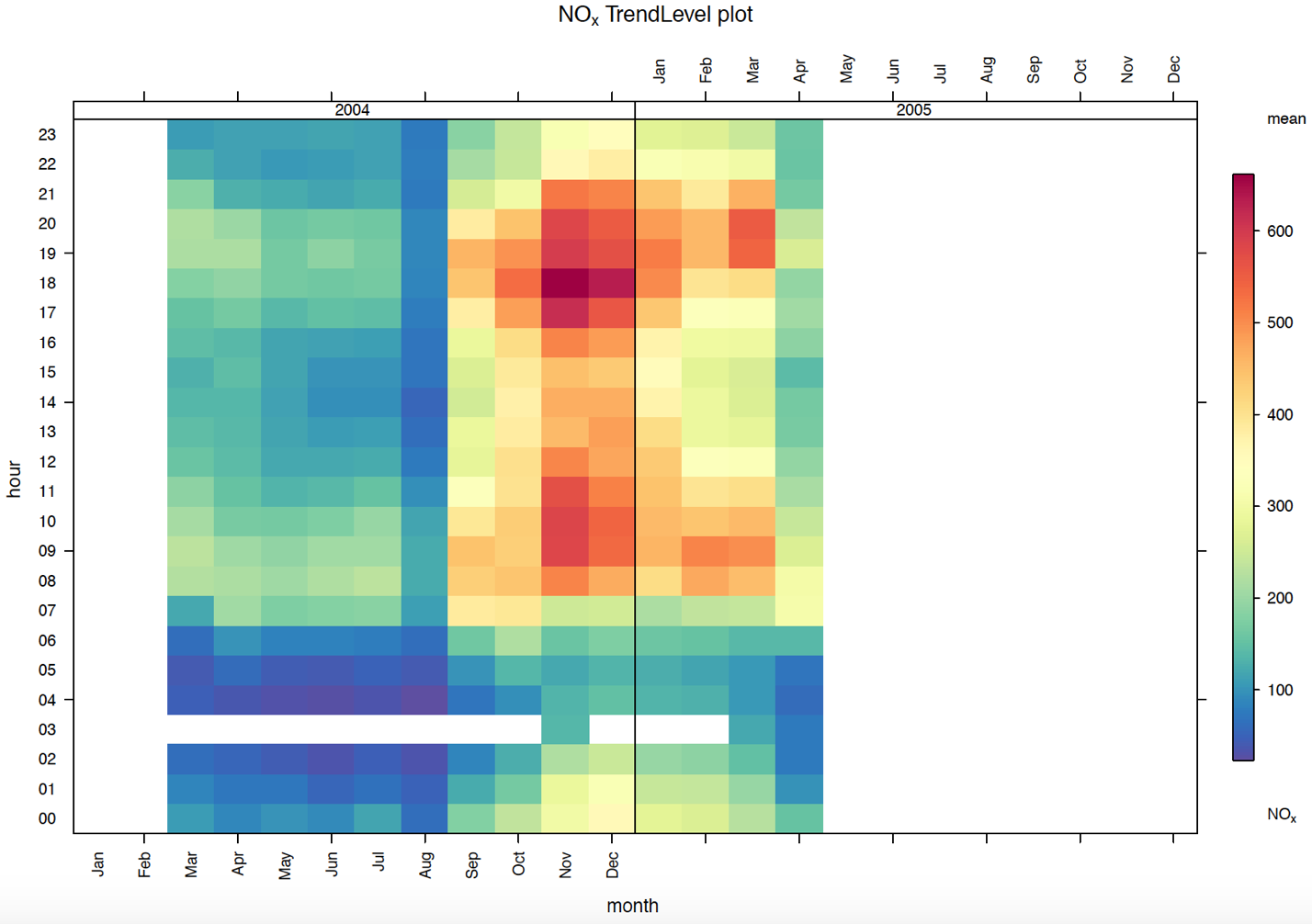
*Carbon Monoxide (CO):*



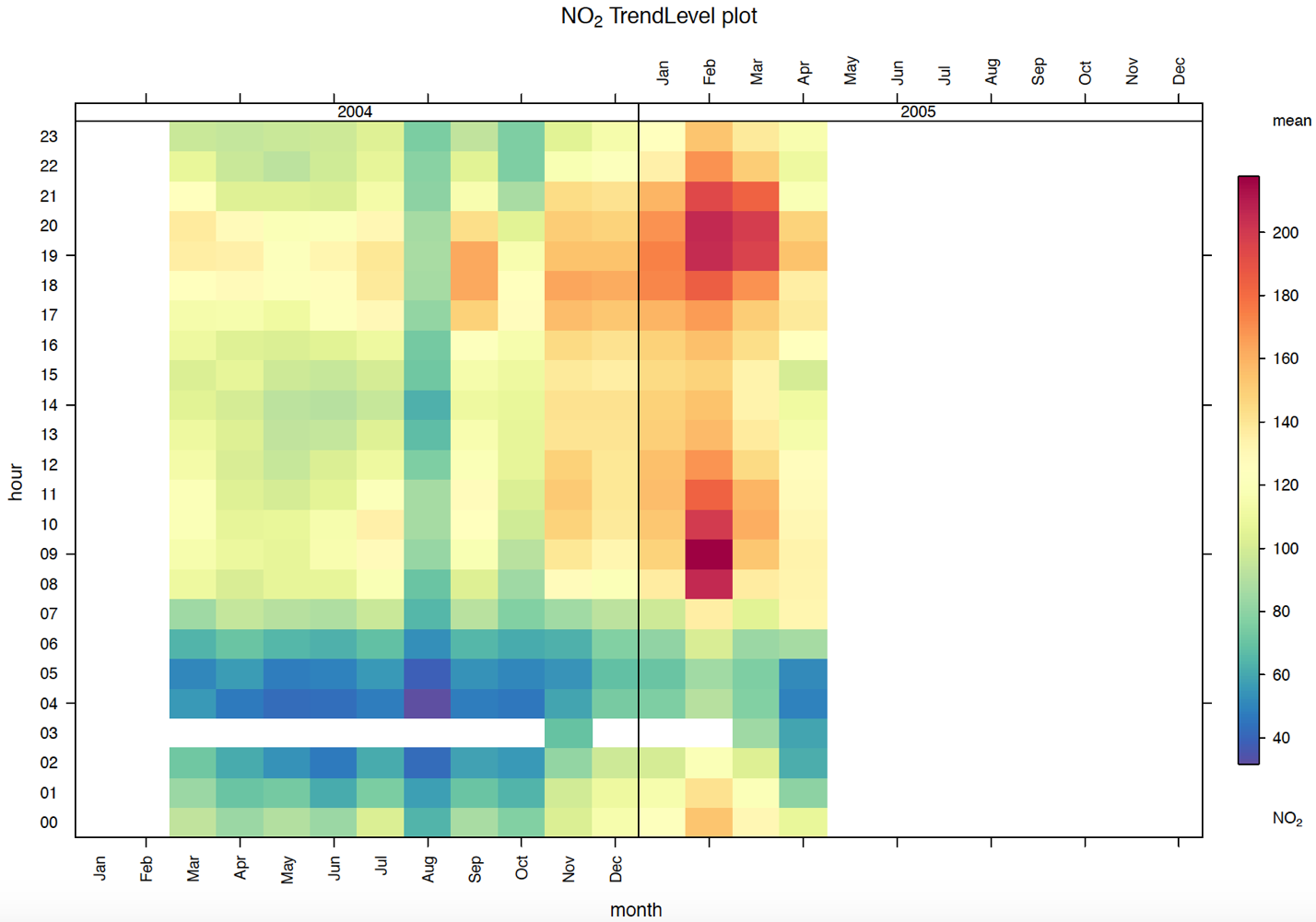
*Benzene (C6H6):*



*Nitrogen Oxides (NOx):*



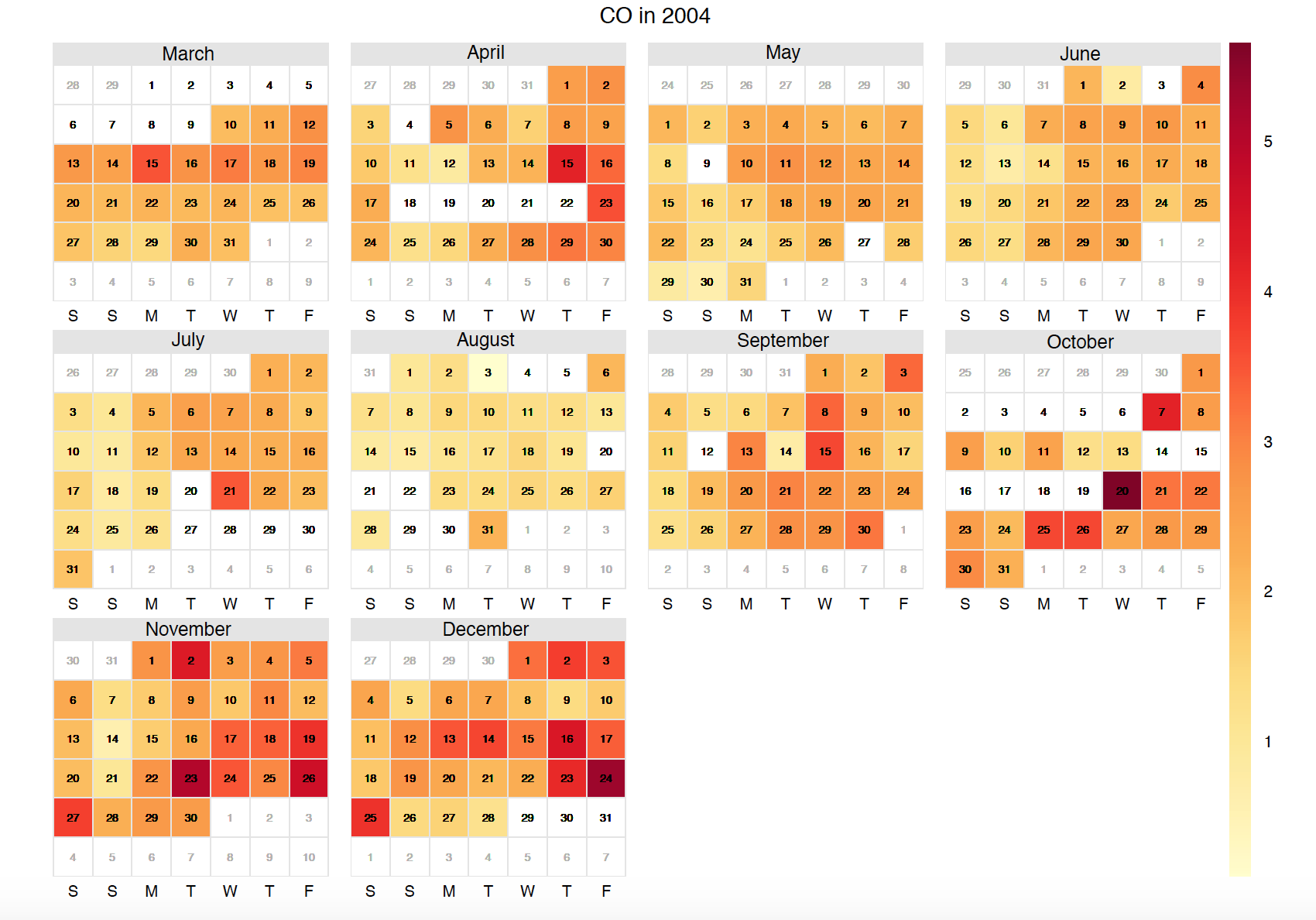
*Nitrogen Dioxide (NO2):*

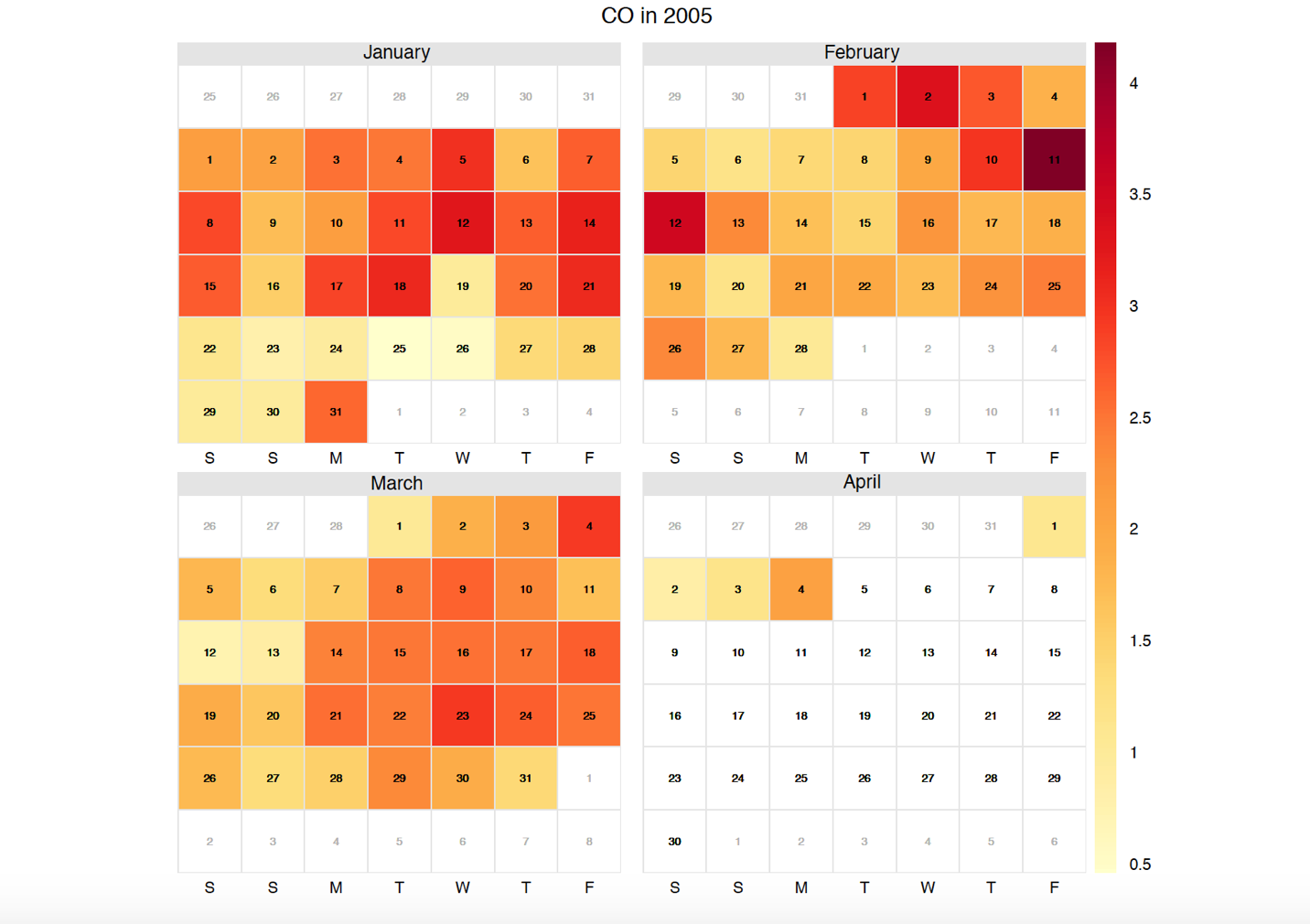


**Calendar Plot:**

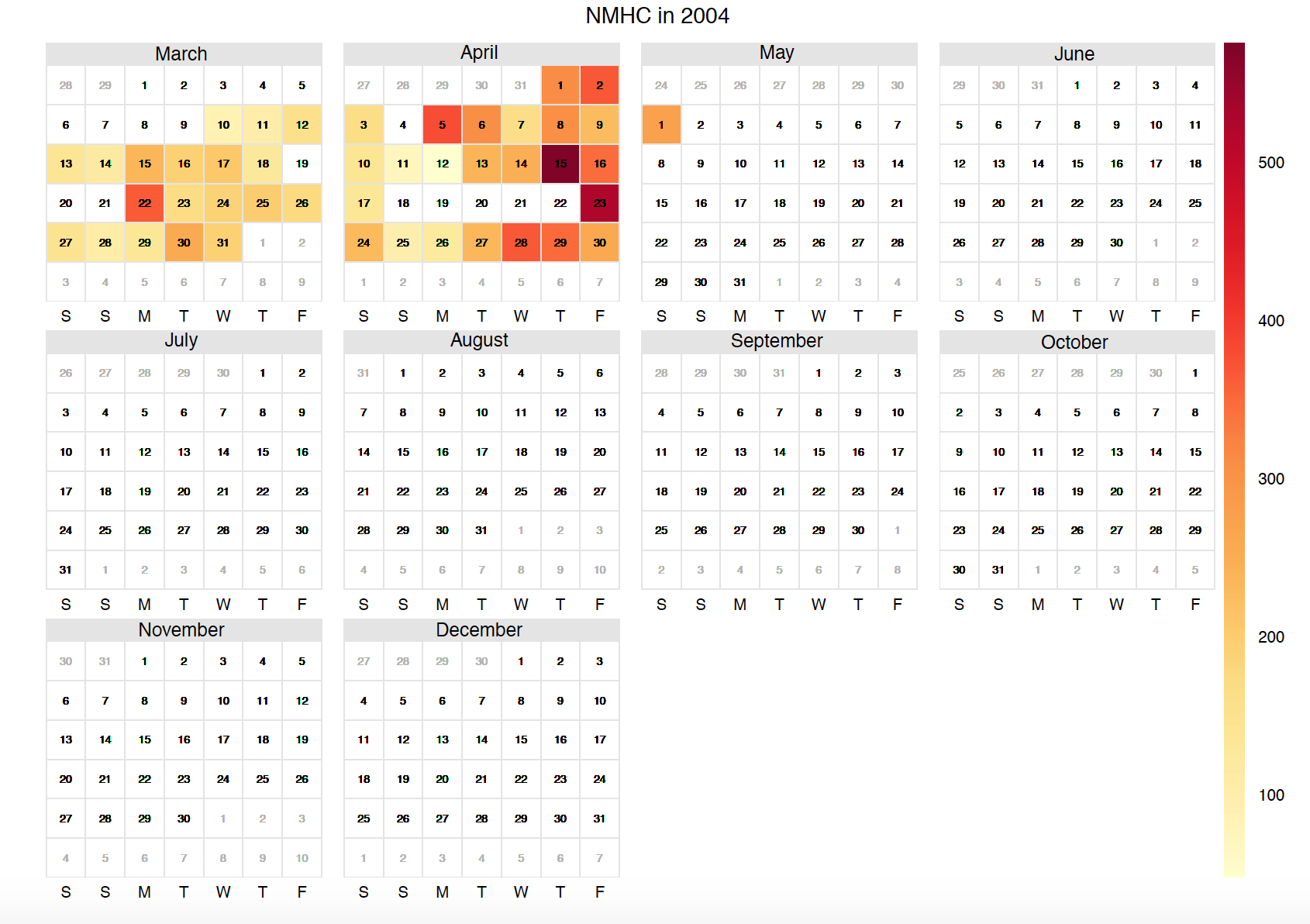
The mean concentration level of a pollutant on each day is plotted on a conventional calendar format. We can compare the level of concentration in the air for each pollutant on a specific day. I have plotted the 2004 calendar year for the Non-Metallic Hydro Carbon data, due to the missing values in NMHC feature. Two calendar plots are plotted for each pollutant in 2004 and 2005 respectively.

*Carbon Monoxide (CO):*

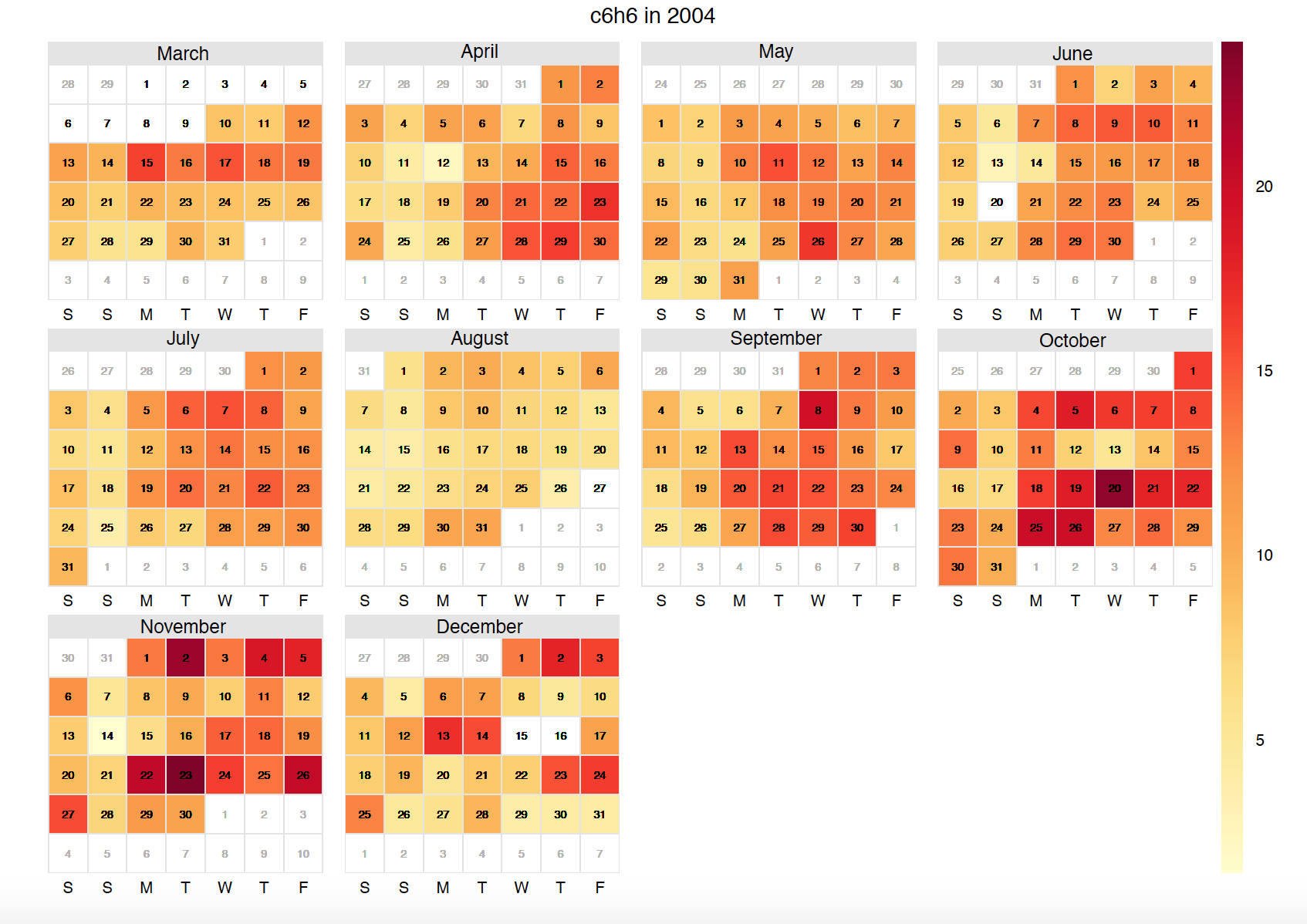


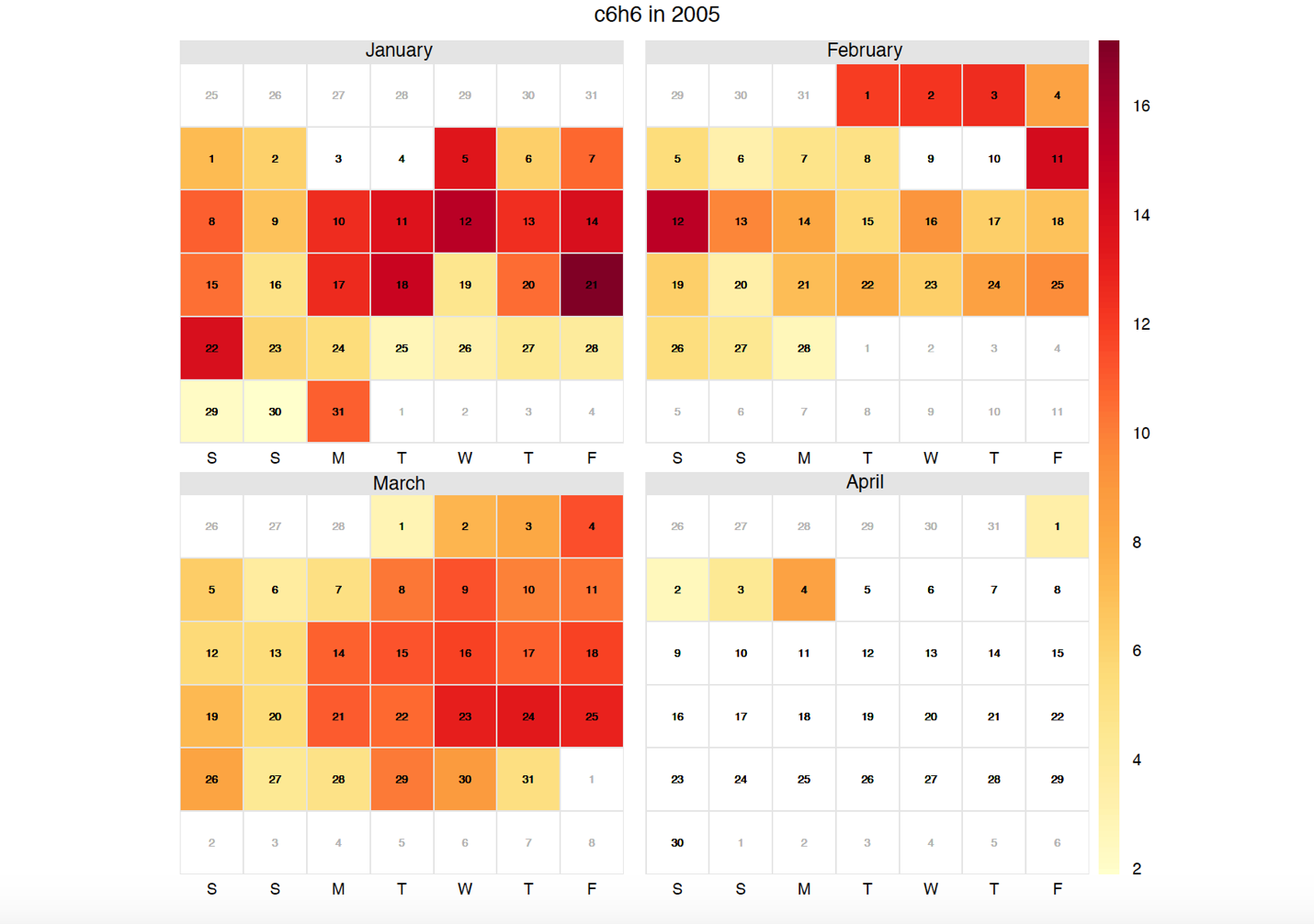


*Non-Metallic Hydro Carbons (NMHC):*

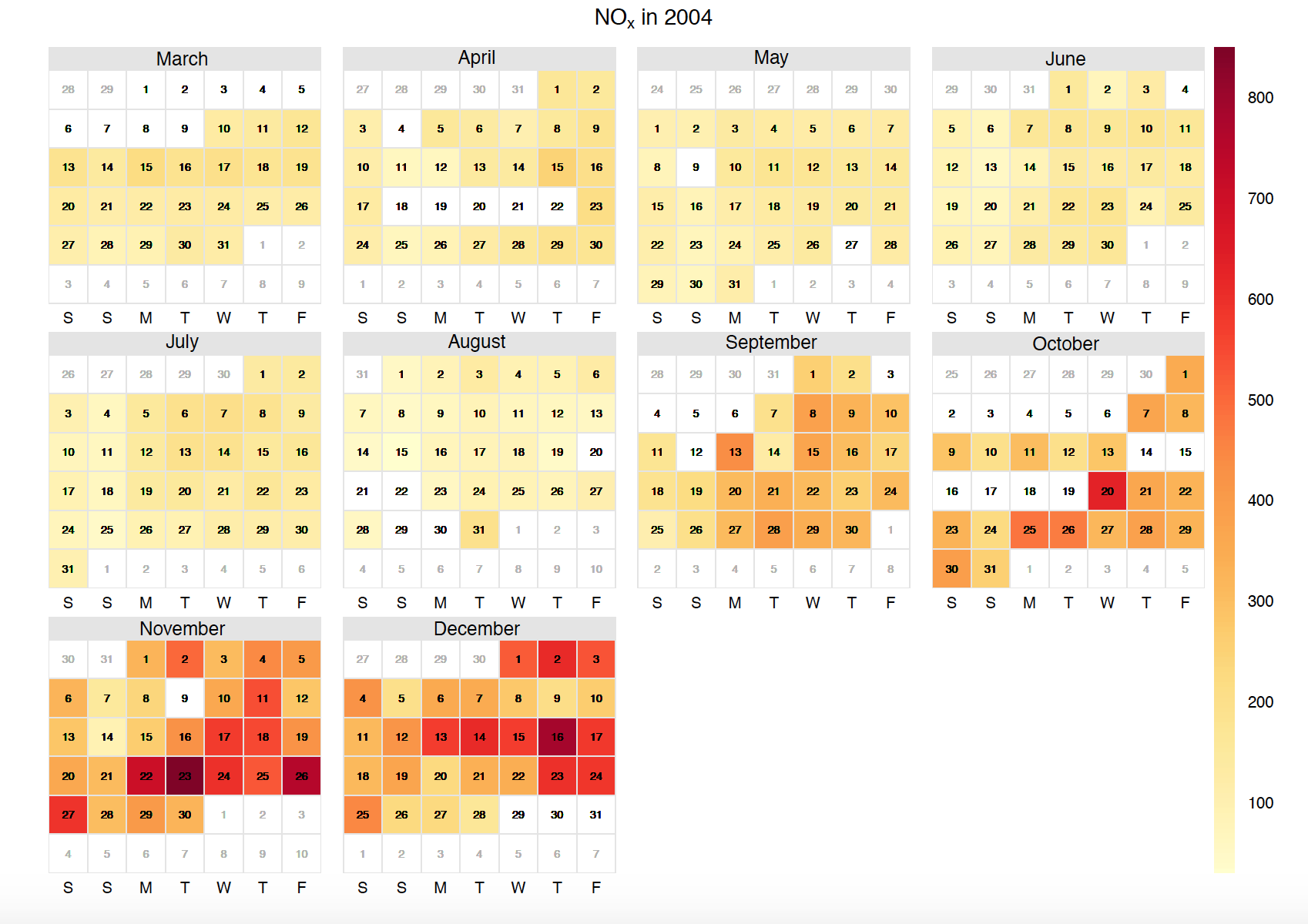


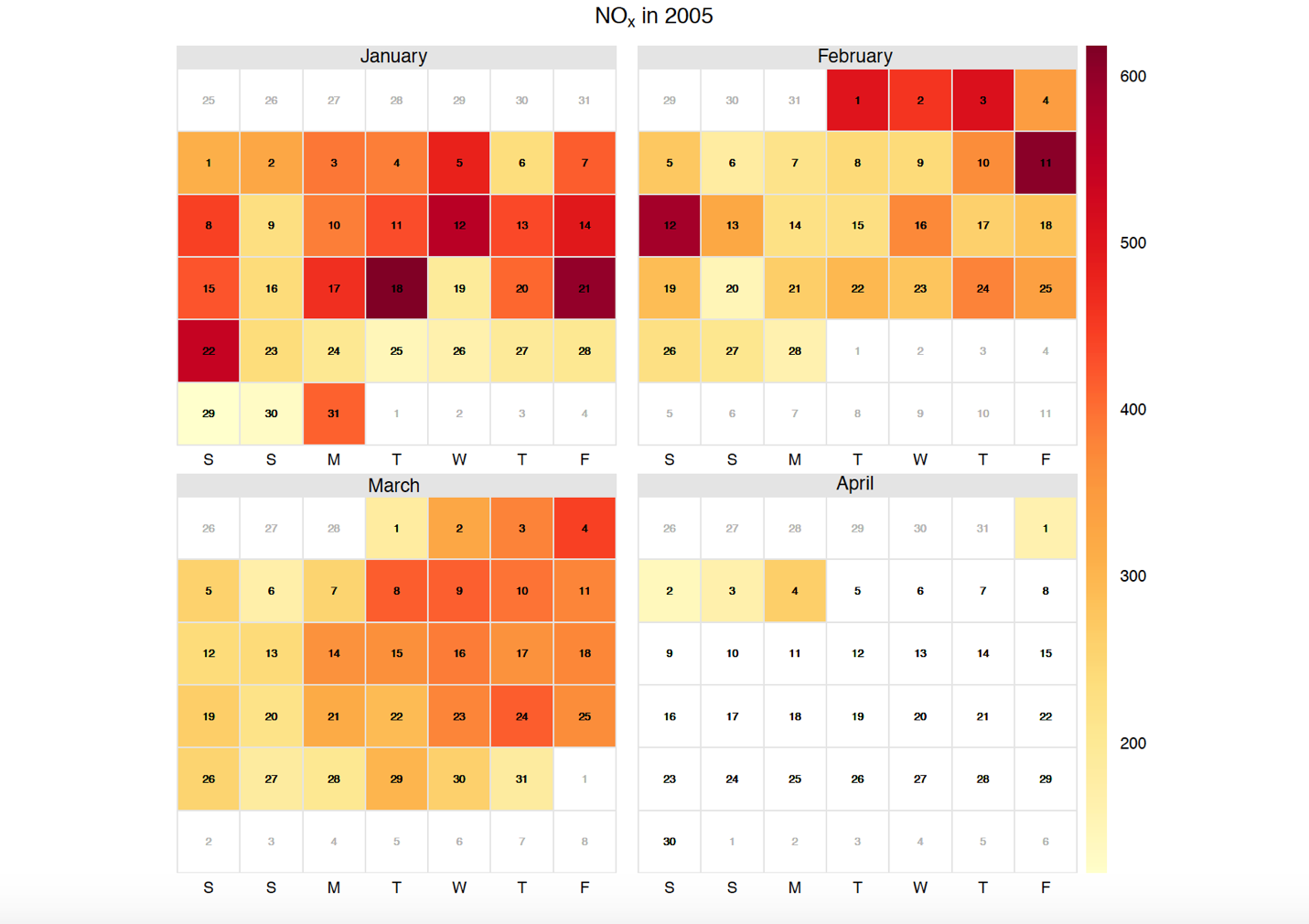
*Benzene (C6H6):*



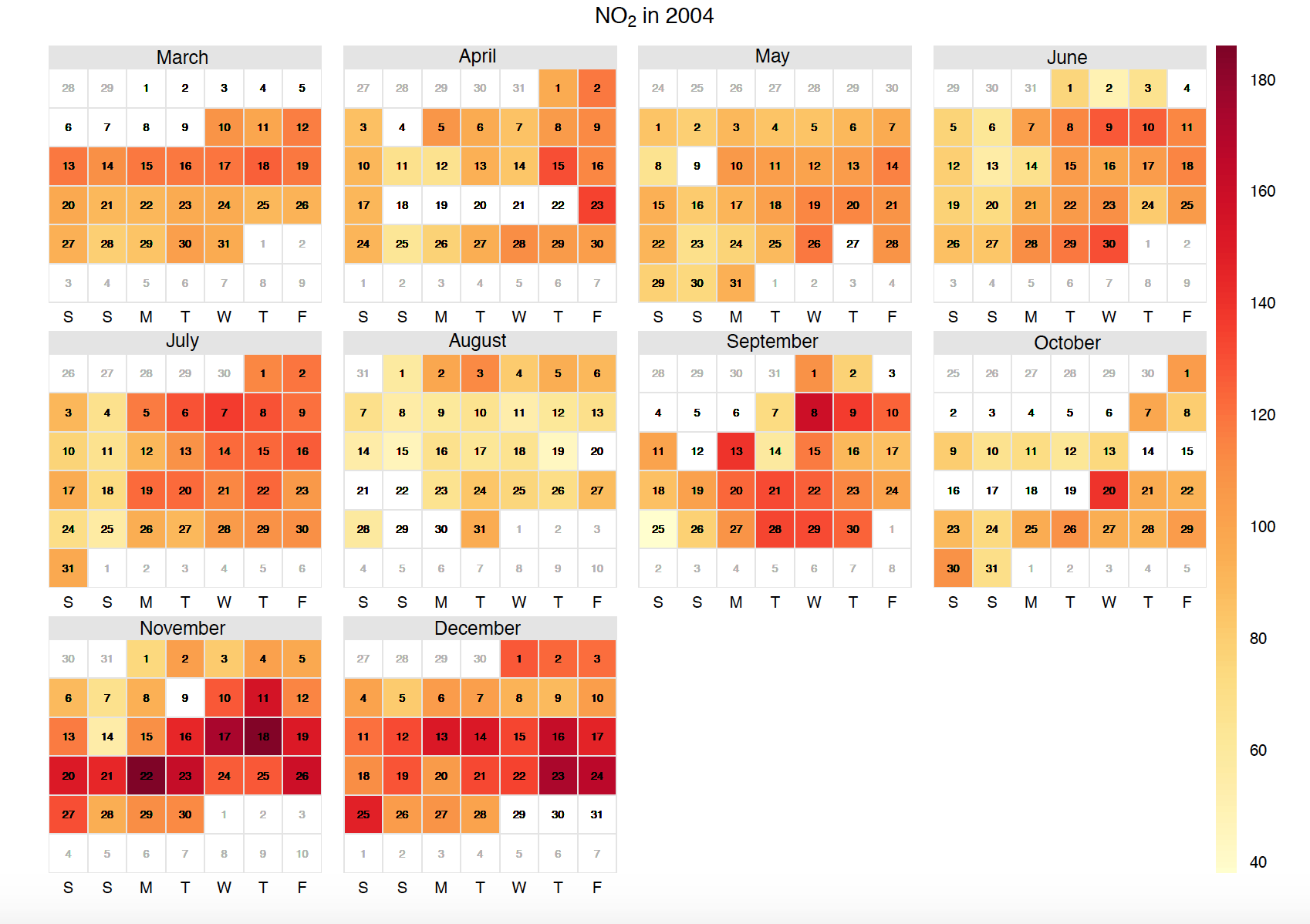


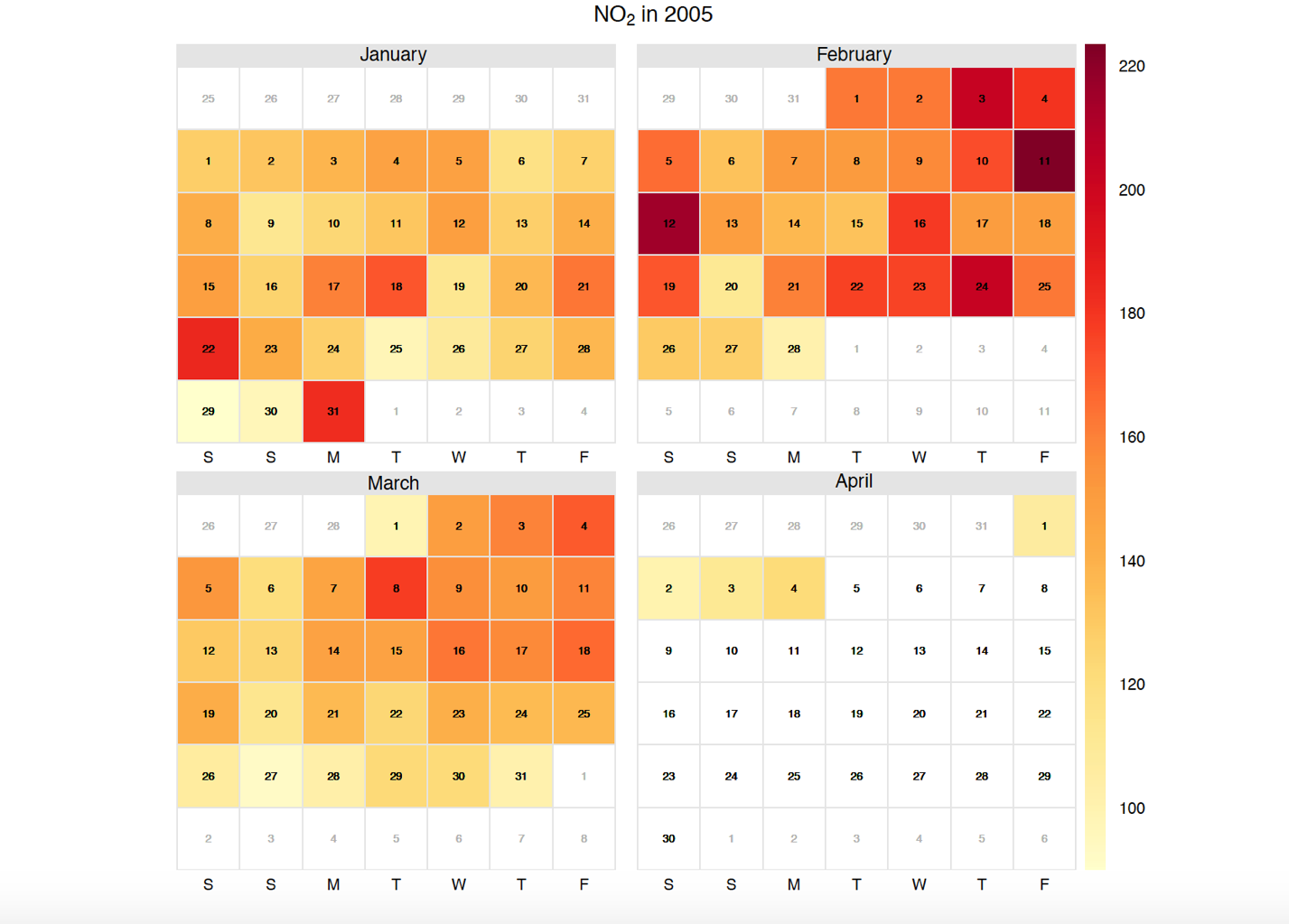
*Nitrogen Oxides (NOx):*





*Nitrogen Dioxide (NO2):*





**Creating a model to predict the Pollutant Concentrations:**

Since there are a lot of missing values in the dataset, we first remove or impute the missing values to maximize our model’s prediction accuracy. For cleaning the data, I have used the data produced by “cmertin” through github:

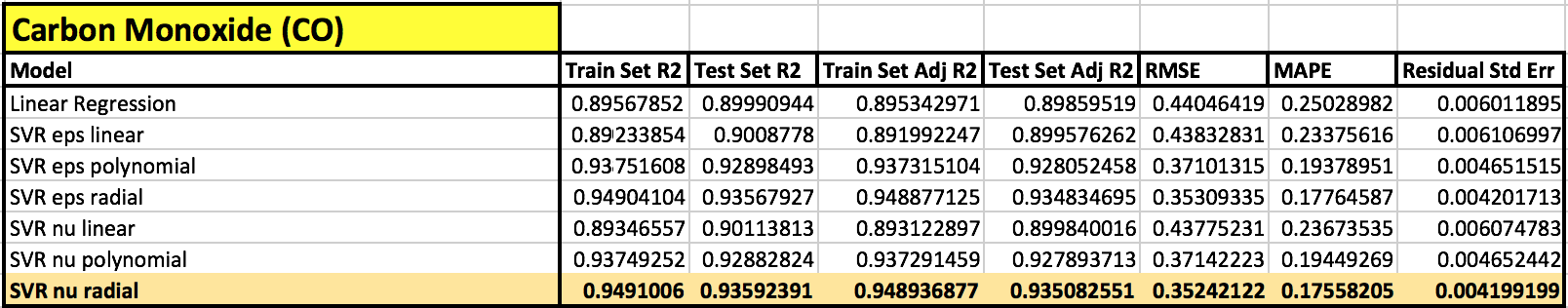
(link: <https://github.com/cmertin/Machine_Learning/tree/master/Air_Quality_Prediction/Data)>.

The data is first parsed and any data points that have more than 2 missing values in a row are removed. Later, he separated the remaining data in terms of with and without missing features. Additionally, he iterated through all the missing variables and performed K-Nearest Neighbors Regression on it, and used the data that had all the features with k = 3. Finally, he calculated the mean of the features and assigned it to the value of the missing feature.

The split ratio for the training set and the test set is 80% and 20% respectively. All the models are trained to predict the ground true values of the pollutants.

**Carbon Monoxide (CO):**

*Predicting the Carbon Monoxide concentration with respect to 5 sensor readings, temperature, relative humidity and absolute humidity.*

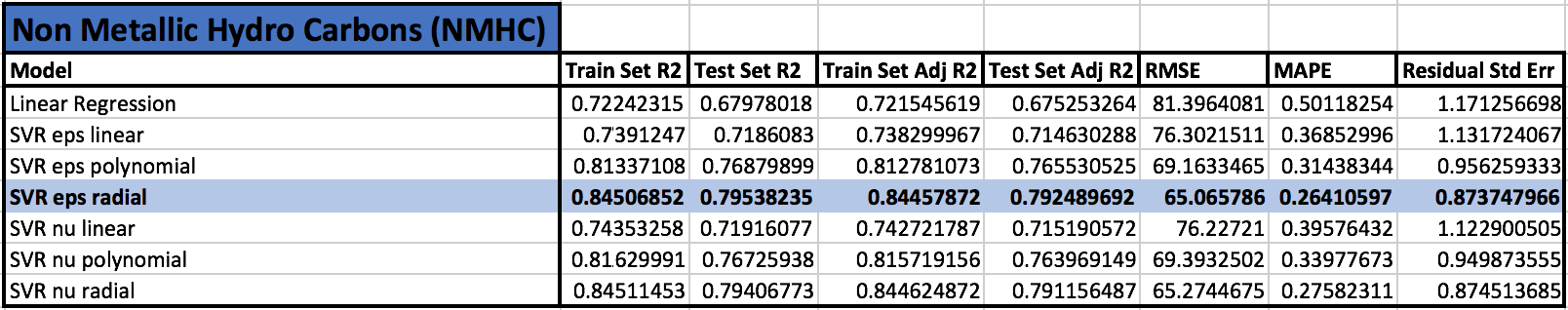


I have trained the dataset targeting the ground true Carbon Monoxide pollutant with Linear Regression and Support Vector Regression (‘eps-linear’, ‘eps-polynomial’, ‘eps-radial’, ‘nu-linear’, ‘nu-polynomial’, ‘nu-radial’).

From the table, according to their performance metrics, we can say that the best model is **Support Vector Regression model** with **‘nu-radial’** kernel**.**

**Non-Metallic Hydro Carbons (NMHC):**

*Predicting the Non-Metallic Hydro Carbons concentration with respect to 5 sensor readings, temperature, relative humidity and absolute humidity.*

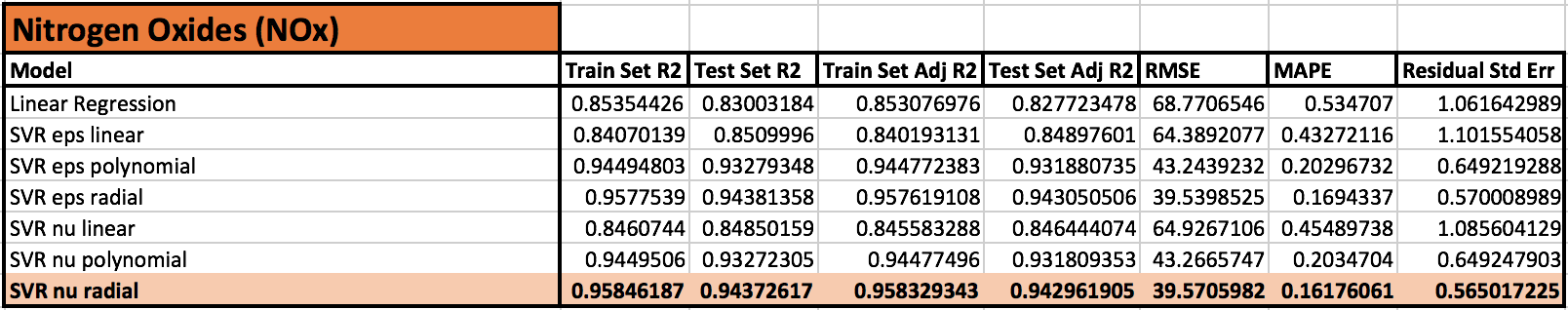


I have trained the dataset targeting the ground true Non-Metallic Hydro Carbons pollutant with Linear Regression and Support Vector Regression (‘eps-linear’, ‘eps-polynomial’, ‘eps-radial’, ‘nu-linear’, ‘nu-polynomial’, ‘nu-radial’).

From the table, according to their performance metrics, we can say that the best model is **Support Vector Regression model** with **‘eps-radial’** kernel**.**

**Nitrogen Oxides (NOx):**

*Predicting the Nitrogen Oxides concentration with respect to 5 sensor readings, temperature, relative humidity and absolute humidity.*

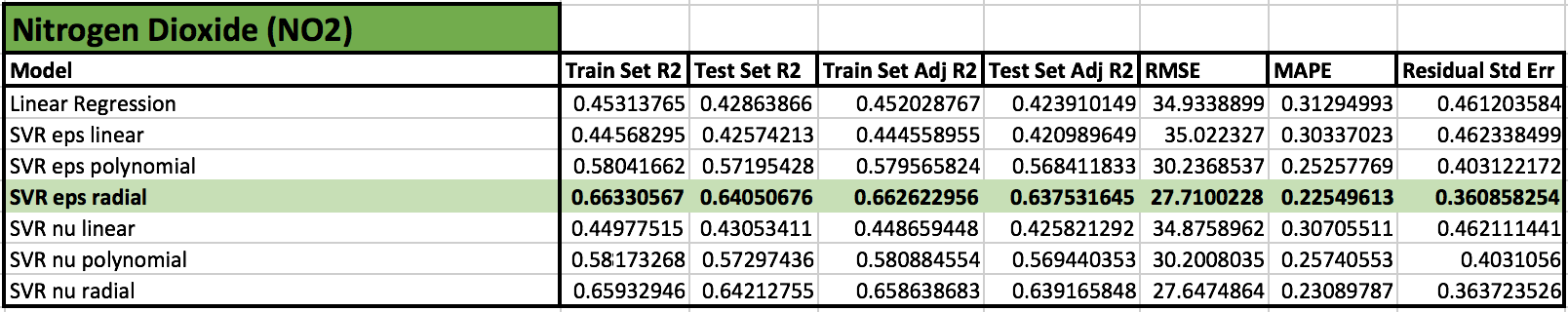


I have trained the dataset targeting the ground true Nitrogen Oxides pollutant with Linear Regression and Support Vector Regression (‘eps-linear’, ‘eps-polynomial’, ‘eps-radial’, ‘nu-linear’, ‘nu-polynomial’, ‘nu-radial’).

From the table, according to their performance metrics, we can say that the best model is **Support Vector Regression model** with **‘nu-radial’** kernel**.**

**Nitrogen Dioxide (NO2):**

*Predicting the Nitrogen Di-Oxide concentration with respect to 5 sensor readings, temperature, relative humidity and absolute humidity.*

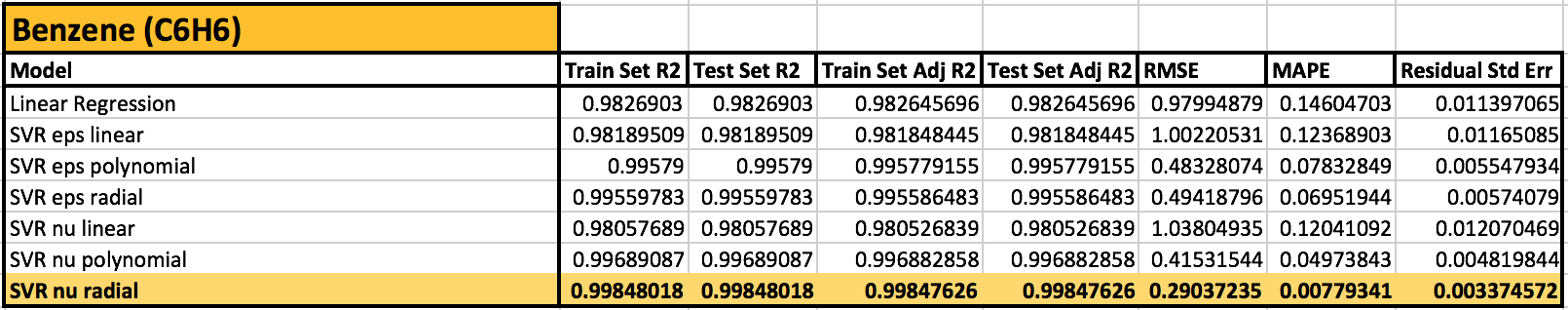


I have trained the dataset targeting the ground true Nitrogen Oxides pollutant with Linear Regression and Support Vector Regression (‘eps-linear’, ‘eps-polynomial’, ‘eps-radial’, ‘nu-linear’, ‘nu-polynomial’, ‘nu-radial’).

From the table, according to their performance metrics, we can say that the best model is **Support Vector Regression model** with **‘eps-radial’** kernel**.**

**Benzene (C6H6):**

*Predicting the Benzene concentration with respect to 5 sensor readings, temperature, relative humidity and absolute humidity.*



I have trained the dataset targeting the ground true Benzene pollutant with Linear Regression and Support Vector Regression (‘eps-linear’, ‘eps-polynomial’, ‘eps-radial’, ‘nu-linear’, ‘nu-polynomial’, ‘nu-radial’).

From the table, according to their performance metrics, we can say that the best model is **Support Vector Regression model** with **‘nu-radial’** kernel**.**

**Conclusion:**

All the models are fitted with linear regression to check the contribution of all the other features on the targeted ground true pollutant. Among all the models SVR models stands out with a better prediction rate and best kernel is either eps-radial or nu-radial for this dataset. Both kernels have almost equal prediction rate with minimal difference. Evidently from all the models, benzene prediction model has the highest prediction rate than any other pollutant prediction model with a residual standard error of **0.0033.**

**References:**

1. Dataset: <https://archive.ics.uci.edu/ml/datasets/Air+quality>
2. Clean Dataset: <https://github.com/cmertin/Machine_Learning/blob/master/Air_Quality_Prediction/Data/AirQuality_clean.csv>