Health Informatio System project

Bettinzoli Nicola

2024-06-18

Introdution

The purpose of this report is to analyse the current state and the evolution of the air quality in the region of Lombardy. During this study i will analize the various concentration of the pollutants in the air and try to identify pattern or trend of the territory and confronting this value with the guidelines and safety value establish by the WHO.

Library used

```
suppressPackageStartupMessages({
  library(dplyr)
  library(ggplot2)
  library(ggmap)
  library(tidyr)
  library(patchwork)
  library(data.table)
  library(ggridges)
  library(geodata)
  library(sf)
  library(osmdata)
  library(gridExtra)
})
```

Datasets used

Dataset air quality sensors

The main dataset utilized is the one containing all the sensor readings, this dataset can be downloaded from the site of the region https://www.dati.lombardia.it/Ambiente/Dati-sensori-aria-dal-2018/g2hp-ar79/about_data. This readings come from the ARPA Lombardy's air quality detection network, this stations automatically read and send back the values of some pollutants of interest. In particular the pollutants continuously monitored are: NOX, SO2, CO, O3, PM10, PM2.5 and Benzene from the 2018 to the 2023.

Let's try to have a look.

```
df <- read.csv("./datasets/sensor_readings.csv")
head(df)
```

```
##   idSensore               Data Valore Stato idOperatore
## 1      5504 01/01/2018 01:00:00   56.9    VA           1
## 2      5504 01/01/2018 02:00:00   53.8    VA           1
## 3      5504 01/01/2018 03:00:00   68.0    VA           1
```

```
## 4      5504 01/01/2018 04:00:00   60.1    VA           1
## 5      5504 01/01/2018 05:00:00   59.7    VA           1
## 6      5504 01/01/2018 06:00:00   56.2    VA           1
```

The feature "idOperatore", probably refers to the organisation/institution that provided the data, which is always set to 1 is not of interest so we just drop it

```
df <- subset(df, select = -c(idOperatore))
```

And the structure

```
str(df)
```

```
## 'data.frame':    17527522 obs. of  4 variables:
##  $ idSensore: int  5504 5504 5504 5504 5504 5504 5504 5504 5504 5504 ...
##  $ Data     : chr  "01/01/2018 01:00:00" "01/01/2018 02:00:00" "01/01/2018 03:00:00" "01/01/2018 04:0
##  $ Valore   : num  56.9 53.8 68 60.1 59.7 56.2 56.8 58.4 61.8 62.4 ...
##  $ Stato    : chr  "VA" "VA" "VA" "VA" ...
```

This dataset consist in more of 17 million records, each of it represents a sensor reading, each reading is described by 4 values: idSensore, Data, Valore, Stato.

The first is simply the ID associate to the sensor, it's an integer value.

The "Date" value indicate the date and the time in whitch the readings was done.

The "Stato" value indicate if the record is valid: in the case of "VA" there is no problem, in the case of "NA" the record is not valid, this can be seen also from the "Valore" feature that, in that case, is set to -9999.

And finally the "Valore" attribute contain the value read by the sensor in that time.

Its important to note that this dataset does not indicate us directly what was the pollutants associated with the readings, for this will be necessary the next dataset that will be utilized to associate the pollutant to the sensor.

Air quality stations dataset

This dataset describe and contains information about the stations and their sensors. All the data can be downloaded from the site of the region https://www.dati.lombardia.it/Ambiente/Stazioni-qualit-dell-aria/ib47-atvt/about_data. The main function of this dataset will be to complement the information seen before.

```
df_stations <- read.csv("./datasets/station_dataset.csv")
head(df_stations)
```

```
##   IdSensore          NomeTipoSensore UnitaMisura Idstazione        NomeStazione
## 1     12691                 Arsenico       ng/m³        560    Varese v.Copelli
## 2      5712                    Ozono       µg/m³        510 Inzago v.le Gramsci
## 3     20488 Particelle sospese PM2.5       µg/m³        564    Erba v. Battisti
## 4     10043            PM10 (SM2005)       µg/m³        687      Ferno v.Di Dio
## 5      6342          Ossidi di Azoto       µg/m³        515    Pero SS Sempione
## 6      6665                    Ozono       µg/m³        565      Cantù v.Meucci
##   Quota Provincia Comune Storico  DataStart   DataStop Utm_Nord UTM_Est
## 1   383        VA Varese       N 01/04/2008             5073728  486035
## 2   138        MI Inzago       S 24/02/2001 01/01/2018  5043030  538012
## 3   279        CO   Erba       N 22/10/2020             5072803  517232
## 4   215        VA  Ferno       N 29/11/2006             5051773  481053
## 5   141        MI   Pero       S 10/12/1986 30/07/2018  5039595  507028
## 6   369        CO  Cantù       N 17/01/2005             5064150  509783
##        lat      lng                      Location
```

```
## 1 45.81697 8.820249  POINT (8.82024911 45.8169745)
## 2 45.53977 9.486897 POINT (9.48689669 45.53976956)
## 3 45.80857 9.221779   POINT (9.2217792 45.8085738)
## 4 45.61925 8.756977 POINT (8.75697656 45.61924753)
## 5 45.50986 9.089974 POINT (9.08997419 45.50985564)
## 6 45.73084 9.125739 POINT (9.12573936 45.73083728)
```

From this dataset we will take some important information like: the pollutants analized by the sensor, the geographical location of it, not only his latitude and longitude, but also the city, the province and the altitude; another useful information will be the unit of measurement of the value read by the sensor.

```
df_stations <- df_stations %>% subset(select = -c(NomeStazione,
                                                  Storico,
                                                  DataStart,
                                                  DataStop,
                                                  Location,
                                                  UTM_Est,
                                                  Utm_Nord))
```

Milan temperature dataset

This dataset can be downloaded from the site of the municipality of Milan https://dati.comune.milano.it/dataset/ds686-rilevazione-temperature-anno-2018 and will be utilized in combination with the previous ones during the study to try to find some connection between the temperature and the emissions. For this we will focus only on the city of Milan in the year of 2018, this because the information about the temperature were not available for all the region.

```
milan_temps <- read.csv("./datasets/milano_temps.csv", sep = ";", na.strings = " ")
```

In this case, for a correct reading of the dataset, specifying the separator and the NA value is necessary.

```
head(milan_temps)
```

```
##        Zone Id.Sensore         Data.Ora Media
## 1  Lambrate       2001 01/01/2018 00:00   3,5
## 2 Zavattari       5920 01/01/2018 00:00   3,2
## 3    Juvara       5909 01/01/2018 00:00     4
## 4    Feltre       8162 01/01/2018 00:00   3,2
## 5     Brera       5897 01/01/2018 00:00   3,6
## 6    Marche       5911 01/01/2018 00:00   3,8
```

```
str(milan_temps)
```

```
## 'data.frame':    52422 obs. of  4 variables:
##  $ Zone      : chr  "Lambrate" "Zavattari" "Juvara" "Feltre" ...
##  $ Id.Sensore: int  2001 5920 5909 8162 5897 5911 2001 5920 5909 8162 ...
##  $ Data.Ora  : chr  "01/01/2018 00:00" "01/01/2018 00:00" "01/01/2018 00:00" "01/01/2018 00:00" ...
##  $ Media     : chr  "3,5" "3,2" "4" "3,2" ...
```

Additional Data

Beside the datasets we will take in consideration also the guidelines established by the WHO for the target value and the safety values for the pollutants, this can be found directly from the publication https://www.who.int/publications/i/item/9789240034228, where the old guidelines (from 15 years ago) has been update thanks to the better knowledge on the effect from the exposure of these pollutants on the human system for prolonged time.

Cleaning and manipuleting the datasets

Renaming features and pollutants

First of all, let us rename all attributes and pollutants for better readability

```
df <- df %>% rename("IdSensor" = "idSensore",
                    "Date" = "Data",
                    "Value" = "Valore",
                    "State" = "Stato")
```

```
df_stations <- df_stations %>% rename("IdSensor" = "IdSensore",
                                      "Type" = "NomeTipoSensore",
                                      "Unit" = "UnitaMisura",
                                      "IdStation" = "Idstazione",
                                      "Altitude" = "Quota",
                                      "Province" = "Provincia",
                                      "Municipality" = "Comune")
```

Then we translate the type of the sensor in the station the dataset

```
df_stations$Type <- df_stations$Type %>% case_match(
  "Monossido di Azoto" ~ "Nitrogen Monoxide",
  "Biossido di Azoto" ~ "Nitrogen Dioxide",
  "Ossidi di Azoto" ~ "Nitrogen Oxides",
  "Biossido di Zolfo" ~ "Sulfur Dioxide",
  "Monossido di Carbonio" ~ "Carbon Monoxide",
  "Ozono" ~ "Ozone",
  "PM10 (SM2005)" ~ "PM10",
  "Particelle sospese PM2.5" ~ "PM2.5",
  "Benzene" ~ "Benzene",
   .default = "NA")
```

Check for NA records

Then we are gonna check how many records inside the dataset are not valid, this are the first to be discarded as they are not useful for our studies and also their mensuration are marked as an error with the value **-9999**.

```
table(df$State, useNA = "always")
```

```
##
##       VA     <NA>
## 16953033   574489
```

```
length(which(df$Value == -9999))
```

```
## [1] 574489
```

As you can see the number of appearances of the value -9999 is equal to the number of NA record.

Since we already have enough data all this record will be simply discharge.

```
df <- drop_na(df)
```

And then we remove the state attribute that is no longer needed.

```
df <- subset(df, select = -c(State))
```

Joining the two dataset

The next step is to merge the two datasets to complete the information of the first and make better consideration on the records.

First of all we check that all the sensors inside the first dataset exist also in the second

```
ids <- df$IdSensor
for (id in ids){
  if (!(id %in% df_stations$IdSensor))
    print("sensor "+id+" not present in the dataset")
}
```

And then we use the **left_join** function to merge the information of the second dataset into the first, in this way we will have all the information needed into one dataset, like the type of the pollutants analyses, the location, etc.

In this study not all the pollutants will be take in consideration but only the one that were monitored continuously by the stations, as indicated in the site of the region.

```
pollutans <- c("Nitrogen Dioxide",
               "Nitrogen Monoxide",
               "Nitrogen Oxides",
               "Sulfur Dioxide",
               "Carbon Monoxide",
               "Ozone",
               "PM10",
               "PM2.5",
               "Benzene")
```

So we just discard the rest.

```
df <- df %>% filter(Type %in% pollutans)
df_stations <- df_stations %>% filter(Type %in% pollutans)
```

Date feature

To be usable the **"Data"** feature has to be converted to a common format, and then read as a date. Also the date and the time will be splitted into two separate features, in this way it will be easier to work with it.

```
dates <- df$Date
```

Convert the string values into date specifying the format, since in the dataset the Italian format is used (dd/mm/yyyy), but the normal is needed.

```
dates <- format(as.POSIXct(dates, format="%d/%m/%Y %H:%M:%S"))
```

For what concerned the time only the hour will be saved, since all the readings are done hourly, minutes and seconds are always at 0.

```
times <- as.integer(format(as.POSIXct(dates), format = "%H"))
```

```
df$Date <- as.Date(dates)

df$Time <- times
```

Finally, for making easier the work during the analysis, to the dataset will be added three features: the year, the month and the day of the week, extracted directly from the date field.

```
df$Year <- df$Date %>% format("%Y")
```

```r
df$Month <- df$Date %>% format("%m")

df$DayOfTheWeek <- df$Date %>% weekdays()
```

Outlier

The next step is to check for outlier inside of the dataset

```r
df %>% group_by(Type) %>% summarize(
  min = min(Value),
  max = max(Value),
  mean = mean(Value),
  quant25 = quantile(Value, probs = 0.25),
  quant50 = quantile(Value, probs = 0.50),
  quant75 = quantile(Value, probs = 0.75)
)
```

```
## # A tibble: 9 x 7
##   Type                min    max   mean quant25 quant50 quant75
##   <chr>             <dbl>  <dbl>  <dbl>   <dbl>   <dbl>   <dbl>
## 1 Benzene             0    200.   0.843     0.2     0.5     1.1
## 2 Carbon Monoxide     0      7.9  0.519     0.3     0.4     0.7
## 3 Nitrogen Dioxide   -0.4  261.  25.9      11.6    21.2    35.7
## 4 Nitrogen Monoxide   0    364.   8.94      0.3     1.3     5.3
## 5 Nitrogen Oxides    -0.3 1747.  46.0      14.5    27.6    56.1
## 6 Ozone               0    370.  51.5      14      46      79.1
## 7 PM10                0    328   28.3      16      24      37
## 8 PM2.5               0    163   20.3      10      16      27
## 9 Sulfur Dioxide     -0.4  353.   2.40      1.1     2       3.2
```

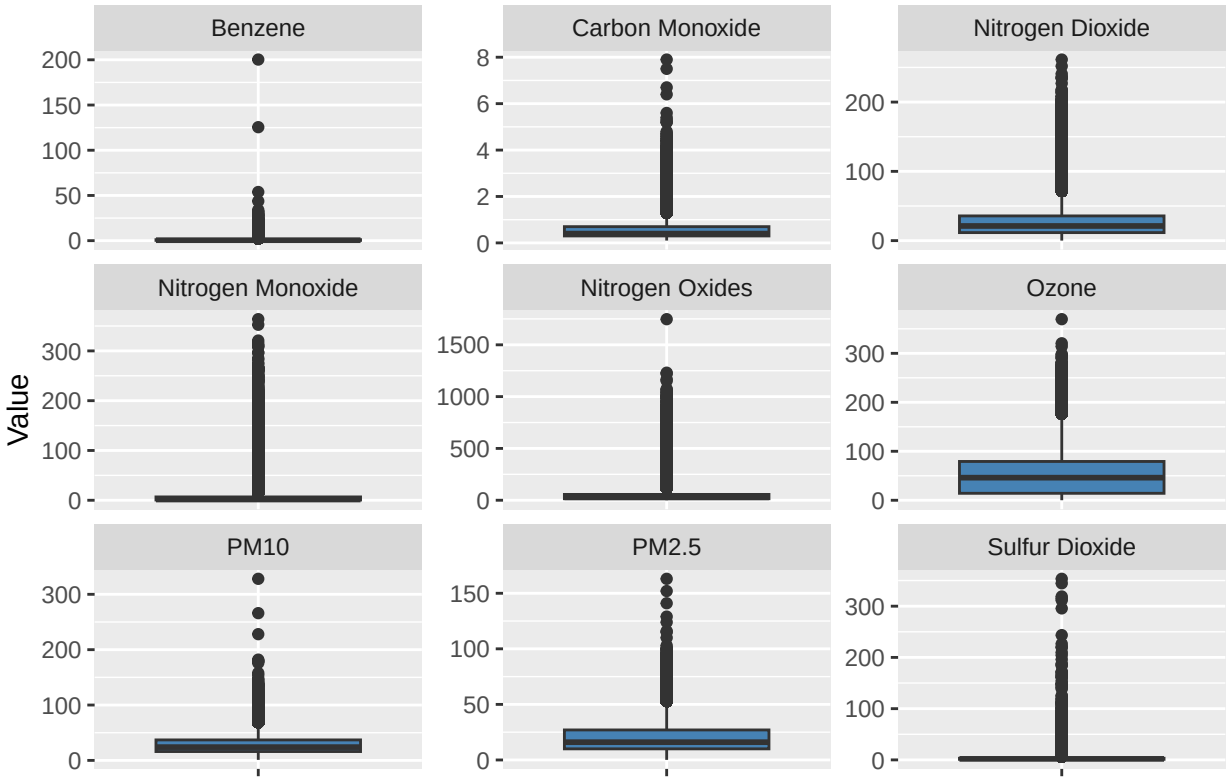
Some value are negative, probably there was some error with the sensor, and there are some strange value way higher than the quant75. For example in the nitrogen oxides the min value is -0.3 and the max is equal to 1747.4, but in we look at the quant75 value is 56.1, way lower that the max observed.

So it might be necessary to discharge some record.

First of all the records with negative value are discarded.

```r
df <- df %>% filter(Value > 0)
```

```r
p <- ggplot(df) +
  aes(x = "", y = Value) +
  geom_boxplot(fill = "steelblue") +
  xlab(" ")

p <- p + facet_wrap(vars(Type),scales = "free_y")
p
```




```
ggsave("./image/outlier_1.png", plot=p, dpi=300)
```

To detect and remove the outlier we use the IQR (Iterquantile ranges), which describe the middle 50% of values when ordered from the lowest to the highest, this value will be used to calculate the threshold to determine which record to keep and discard. The IQR is calculated as the difference between the the third (quant75) and the first (quant25) quartile.

$$IQR = Q_3 - Q_1$$

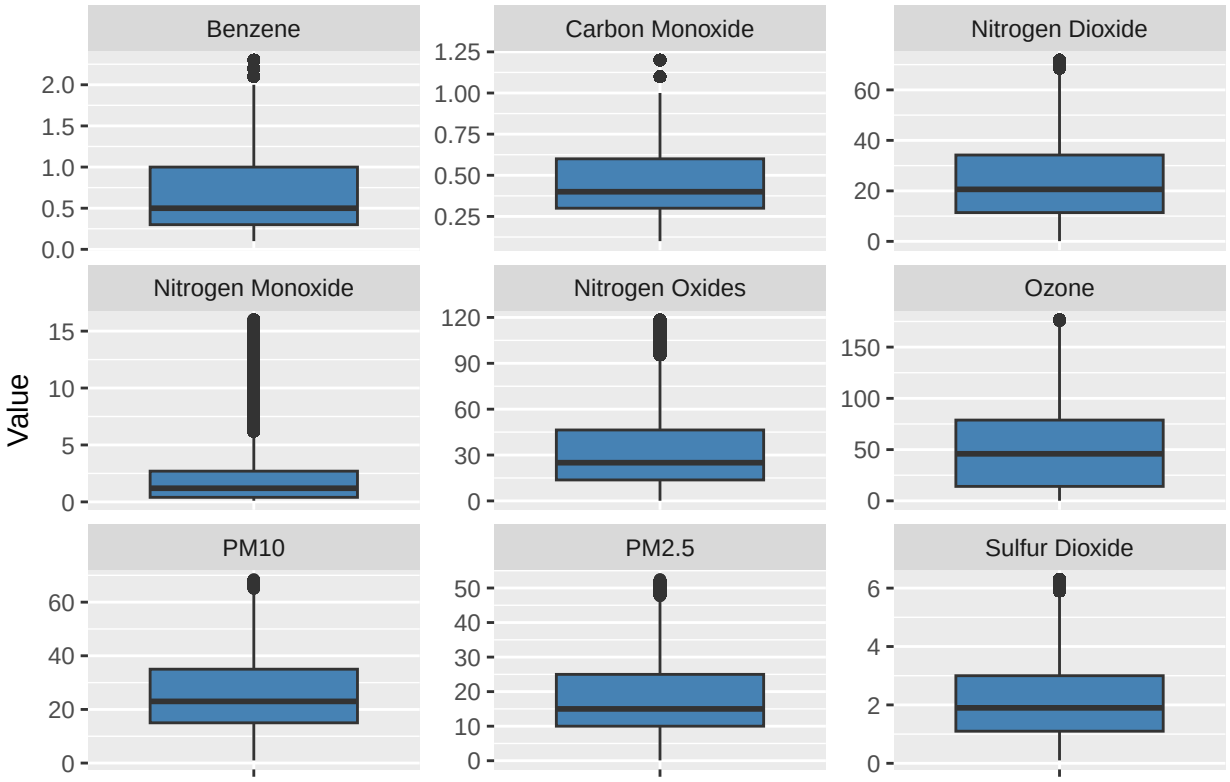
Then this value is utilized the determine the threshold by multiplying it with a constant, for example in this case for 1.5.

```
remove_outlier <- function(df){

  category <- unique(df$Type)

  for (c in category){
    values <- df$Value[df$Type == c]

    quantile25 <- quantile(values, probs = 0.25)


    quantile75 <- quantile(values, probs = 0.75)

    IQR <- quantile75 - quantile25
```

```
    df <- df %>% filter(!(Type == c & (Value > quantile75 + (IQR * 1.5))))
  }
  return(df)
}
```

```
df <- remove_outlier(df)
```

```
p <- ggplot(df) +
  aes(x = "", y = Value) +
  geom_boxplot(fill = "steelblue") +
  xlab("")

p <- p + facet_wrap(vars(Type),scales = "free_y")
p
```




```
ggsave("./image/outlier_2.png", plot=p, dpi=300)
```

We can also have a look at the new min/max/mean values after the removal of the outlier.

```
## # A tibble: 9 x 7
##   Type                min   max   mean quant25 quant50 quant75
##   <chr>             <dbl> <dbl>  <dbl>   <dbl>   <dbl>   <dbl>
## 1 Benzene             0.1   2.3  0.674     0.3     0.5     1
## 2 Carbon Monoxide     0.1   1.2  0.481     0.3     0.4     0.6
## 3 Nitrogen Dioxide    0.1  71.8 24.3      11.4    20.6    34.2
## 4 Nitrogen Monoxide   0.1  16    2.46      0.4     1.2     2.7
## 5 Nitrogen Oxides     0.1 118.  33.7      13.7    25      46.4
```

```
## 6 Ozone               0.1 177.  51.2      14.1    45.9    78.8
## 7 PM10                1    68.5 26.3      15      23      35
## 8 PM2.5               0.1  52.5 18.6      10      15      25
## 9 Sulfur Dioxide      0.1   6.3  2.18      1.1     1.9     3
```

As we can see now the max values are less distant from the mean, and they are not to far away from the quant75.

Siamo partiti con più di 17 milioni di rilevazioni, adesso finita la pulizia siamo a circa 15 milioni.

```
nrow(df)
```

```
## [1] 15540280
```

Cleanng the tempertature datastes

Finally, with regard to the temperature dataset, some modifications were necessary: first of all it was necessary, during the reading for a correct interpretation, to specify the separator and the null term, after which it was also necessary to replace the separator between the integer and the decimals with the dot, for a correct conversion then to float type data (necessary for subsequent operations), and finally, ,like with the previous dataset, we convert the Data.ora attribute as a date type.

```
milan_temps <- drop_na(milan_temps)
milan_temps$Media <- gsub(",", ".", milan_temps$Media)
milan_temps$Media <- as.numeric(milan_temps$Media)
milan_temps$Data.Ora <- as.Date(milan_temps$Data.Ora, format = "%d/%m/%Y %H:%M")
```

Of this dataset only the necessary information such as date and temperature were kept, of these then the daily average was taken.

```
milan_temps <- milan_temps %>% group_by(Data.Ora) %>% summarise(Media = mean(Media))
```

Analyses

Distribution

First we have gonna look on how the stations are distributed in the territory, the source stated that the sensors and the stations are distributed based on the density of population.

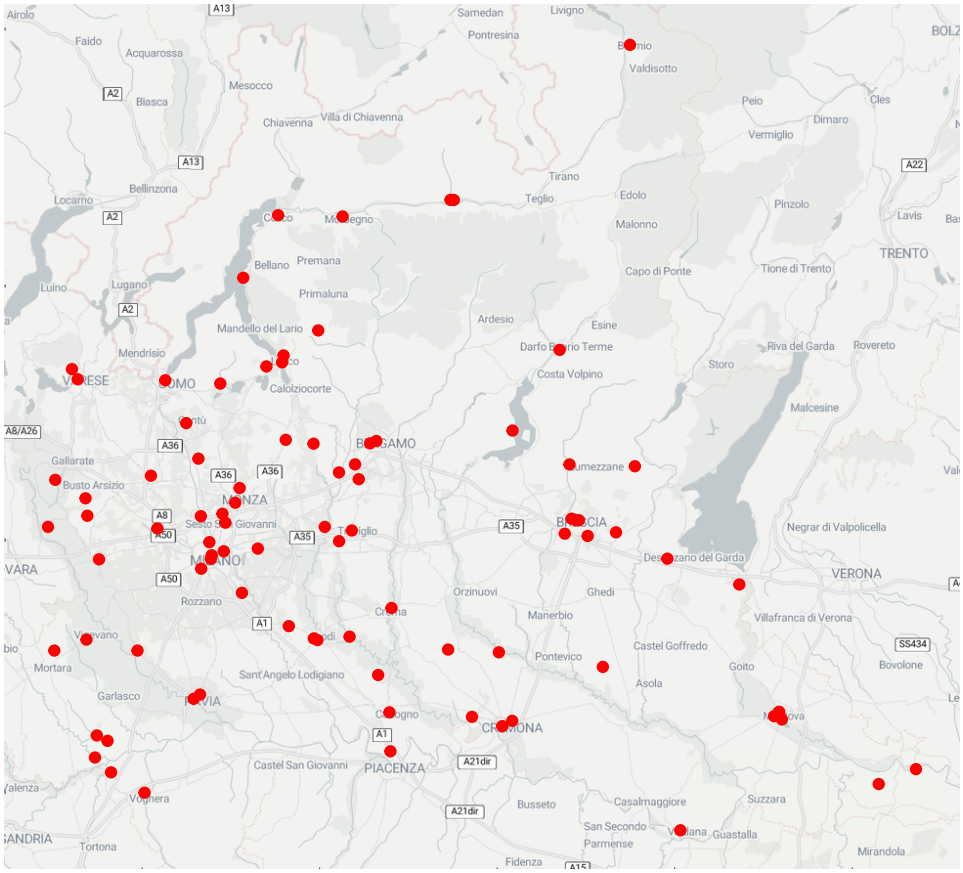
```
station <- df %>%
  subset(select = c(IdStation,lat,lng,Altitude)) %>%
  group_by(IdStation) %>%
  summarise(lat = mean(lat),
            lng = mean(lng),
            Altitude = mean(Altitude))
```

To retrieve the map it will be used Stadia, because it can be used for free and it doesn't need to enter a credit card to obtain an api key.

```
api_key <- readLines("stadia_key.txt")
register_stadiamaps(api_key, write = FALSE)
```

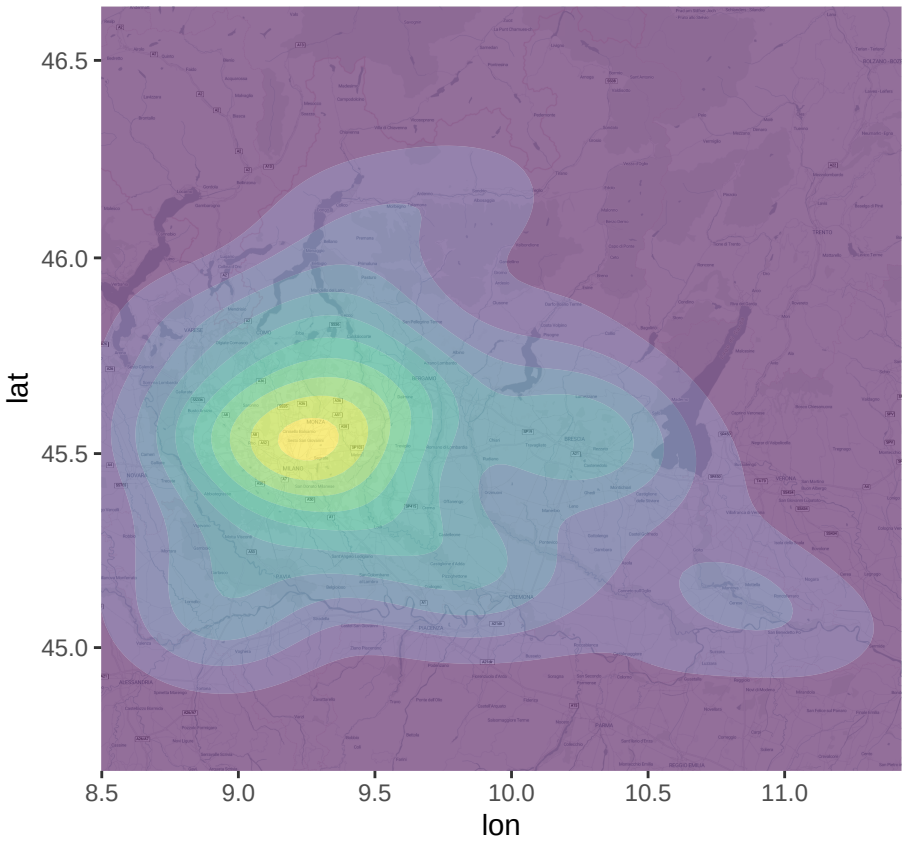
Then the station are plotted on the map

```
qmplot(x=lng,y=lat,data = station, source = "stadia",maptype = "alidade_smooth", color = I("red"))
```




```
bb <- getbb("Lombardia")
map <- get_stadiamap(bbox = c(bb[1,1],bb[2,1],bb[1,2],bb[2,2]), maptype = "alidade_smooth")
```

```
p <- ggmap(map) +
  geom_density_2d_filled(data = station,
                         aes(x=lng, y=lat, alpha = 0.3)) +
  theme(legend.position = "none")
p
```

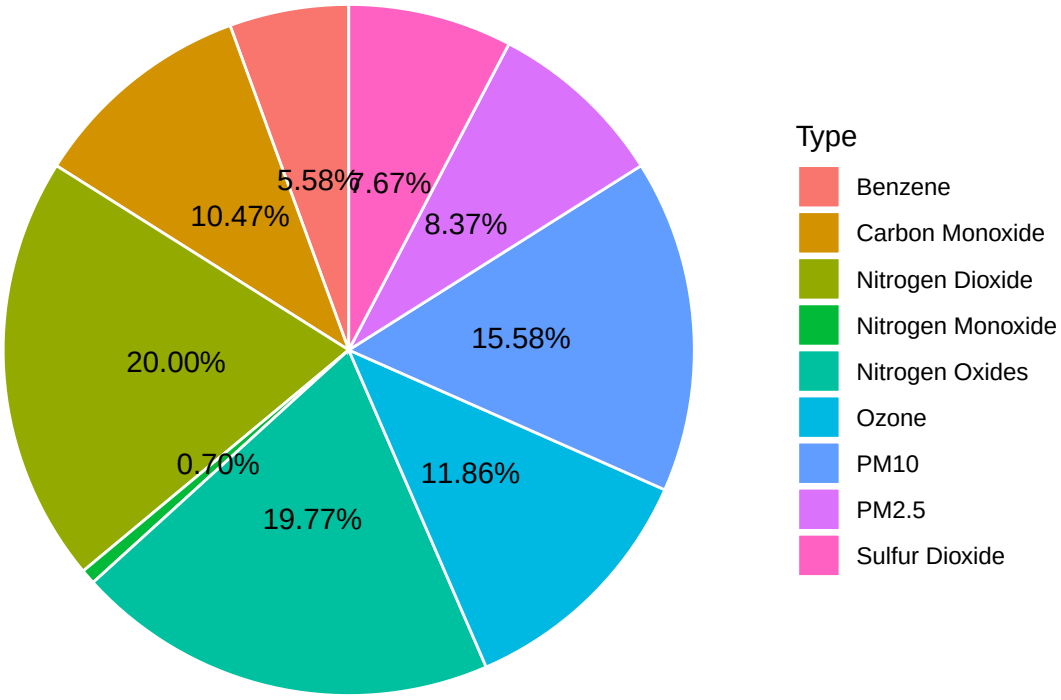



```
ggsave("./image/stations_desnity.png", plot=p, width=13, height=7, dpi=300)
```

As we can see, like stated by the source, the stations are located near towns and more in general close to more populated areas, like Milan, Brescia, Bergamo, etc.

Then we are gonna check the number of sensors for each pollutant

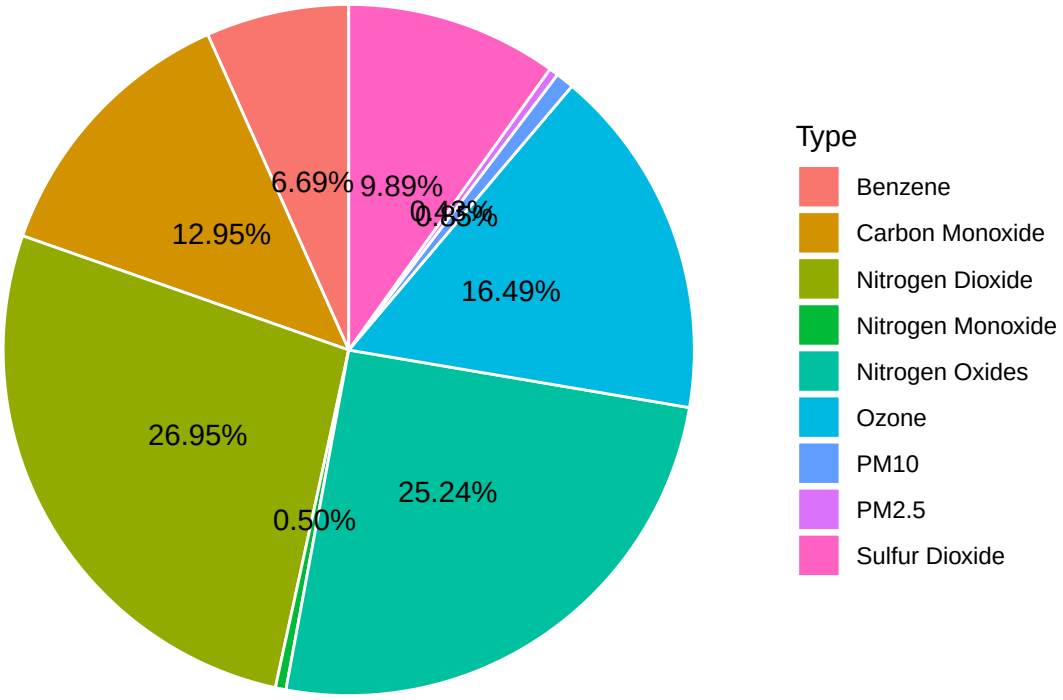
```
sensor_count <- df_stations %>%
  group_by(Type) %>%
  filter(Type %in% pollutans & IdSensor %in% df$IdSensor) %>%
  summarise(n = n())

total <- sum(sensor_count$n)

sensor_count$perc <- sensor_count$n / total
sensor_count$perc <- sprintf("%1.2f%%", 100*sensor_count$perc)


ggplot(sensor_count, aes(x="", y=n, fill=Type)) +
  geom_bar(stat="identity", width=1, color="white") +
  coord_polar("y", start=0) +
  geom_text(aes(label = perc), position = position_stack(vjust=0.5)) +
  theme_void()
```



In the graph it's possible to see how the majority the number of sensor are equally distributed all ranging from the 15% to the 6%. The big exceptions are the Nitrogen Dioxide/Oxides that stands above and the Nitrogen Monoxide that has far less sensor.

Then the number of readings for each pollutant, in this case we will expect to have less readings for the PM10 and the PM2.5, this because for both there's only one record a day that represents the daily mean, the others have it hourly.

```
raeadings_count <- df %>%
 group_by(Type) %>%
 summarise(n = n())
total <- sum(raeadings_count$n)
raeadings_count$perc <- raeadings_count$n / total
raeadings_count$perc <- sprintf("%1.2f%%", 100*raeadings_count$perc)
p <- ggplot(raeadings_count, aes(x="", y=n, fill=Type)) +
 geom_bar(stat="identity", width=1, color="white") +
 coord_polar("y", start=0) +
 geom_text(aes(label = perc), position = position_stack(vjust=0.5)) +
 theme_void()
p
```




```
ggsave("./image/readings_count_pie.png", plot=p, width=7.5, height=5, dpi=150)
```

Like stated before, PM10 and PM2.5 have less rilevations, due to the frequency of feedback different from the others. The other one that is behind is the Nitrogen Monoxide, and this is due to the lower number of sensors, seen in the previous graph.

The remaining are not too far from each other, with the only exception of Nitrogen Dioxide/Oxides that are a bit higher at 25% each, due to the higher number of sensors.

Studing the averages

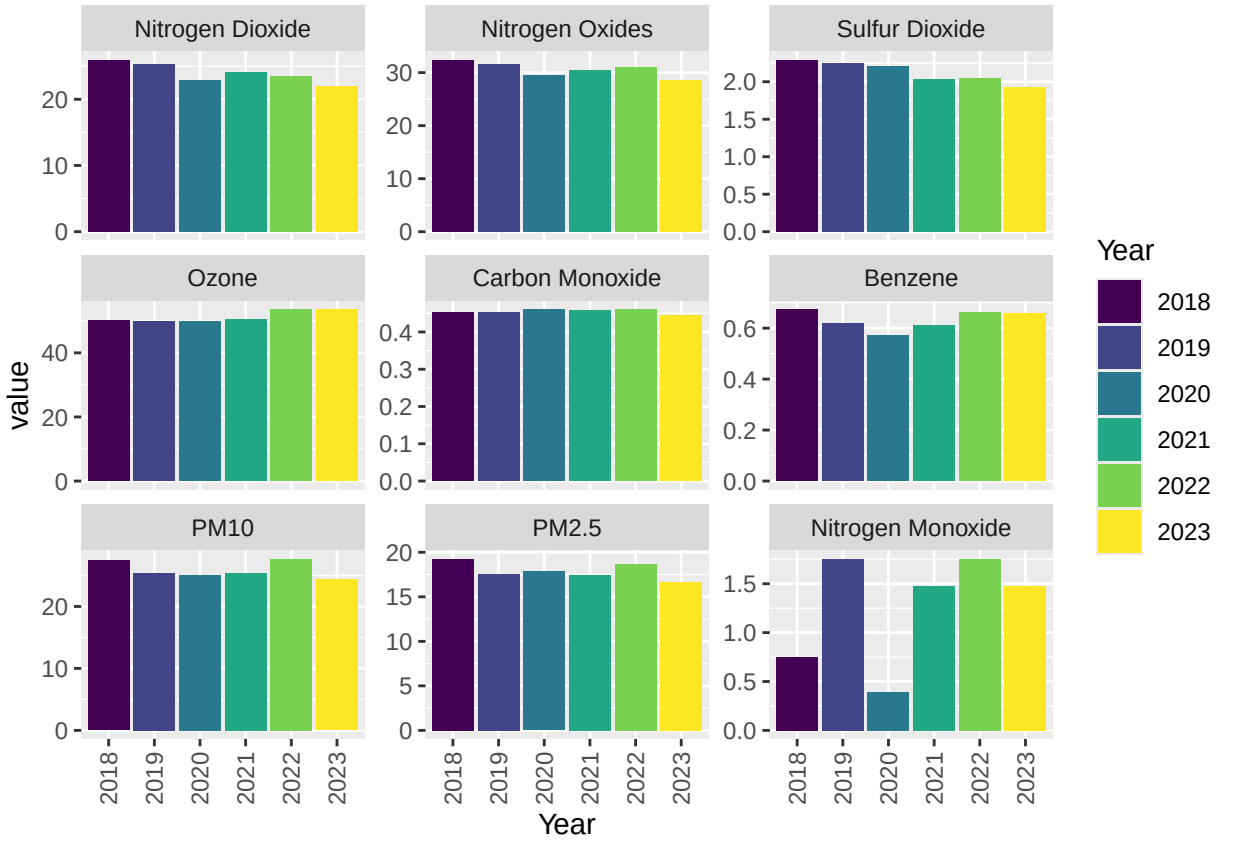
Now we are gonna study how the concentration of each pollutants varies from different points of view, the code utilized for each of it is very similar, so it will be displayed only in the first experiment.

The first in consideration is how the levels varies between the years, from the 2018 to the 2023

```
annual_averages <- df %>%
  subset(select = c(Type,Value,Year)) %>%
  filter(Year < 2024) %>%
  pivot_wider(names_from = Type, values_from = Value, values_fn = mean)
```

```
years <- 2018:2023
annual_averages <- setDT(annual_averages)
mtab = melt(annual_averages, id.vars="Year")

mtab$Year = factor(mtab$Year, levels=c(years))

p <- ggplot(data=mtab, aes(x=Year, y=value, fill=Year)) +
```

```
  geom_bar(stat="identity") +
  scale_fill_viridis_d() +
  facet_wrap(vars(variable),scales = "free_y") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))

p
```




```
ggsave("./image/years.png", plot=p, width=7.5, height=5, dpi=150)
```

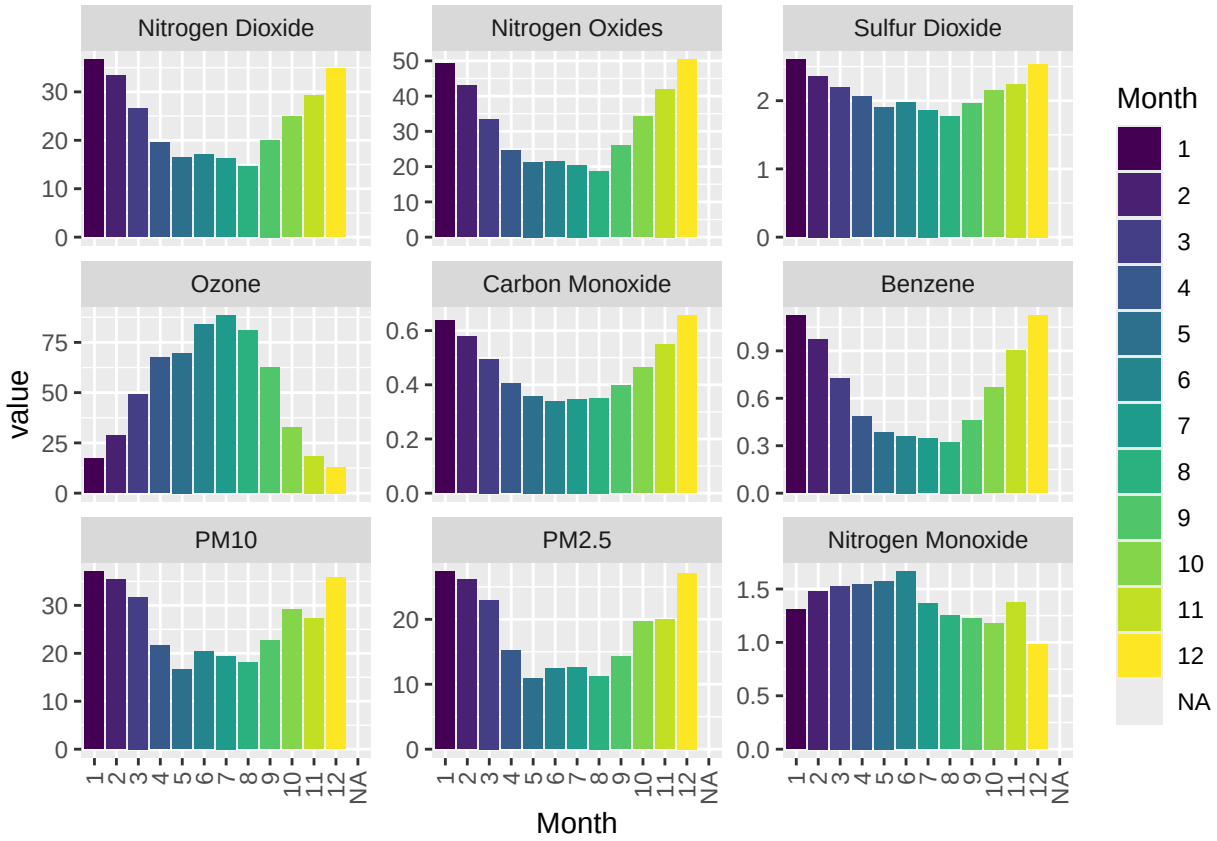
From the graph it's possible to see that there were some strange behavior with the nitrogen monoxide, that has two sudden drops in the 2018 and in the 2020. Probably this is not due to the pandemic, because it is not only related to the 2020, but more likely caused by some problem with the stations or the sensors.

For what concern the other pollutants are in slight decrease, like the nitrogen dioxide or the sulfur dioxides, others, for example the PM10 and PM2.5, remain fairly stationary, with particular peaks in the 2022.

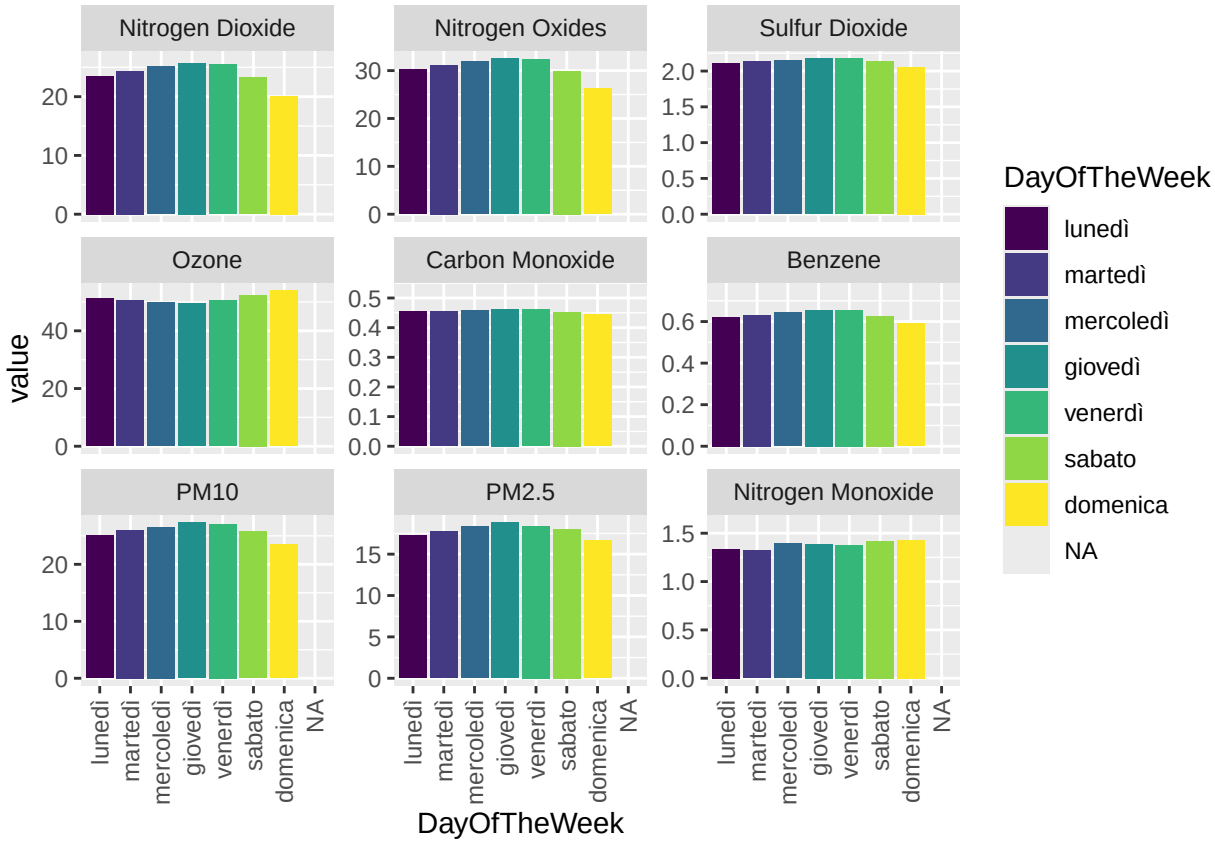
It's interesting to see how the pandemic didn't impact the level for none of the pollutants.

There was also some data from the 2024, but they are related only for one day, this because the rilevations reach as far as the 2024-01-01 so for this test they got excluded.

Secondly for each month, in this case all the data were used.

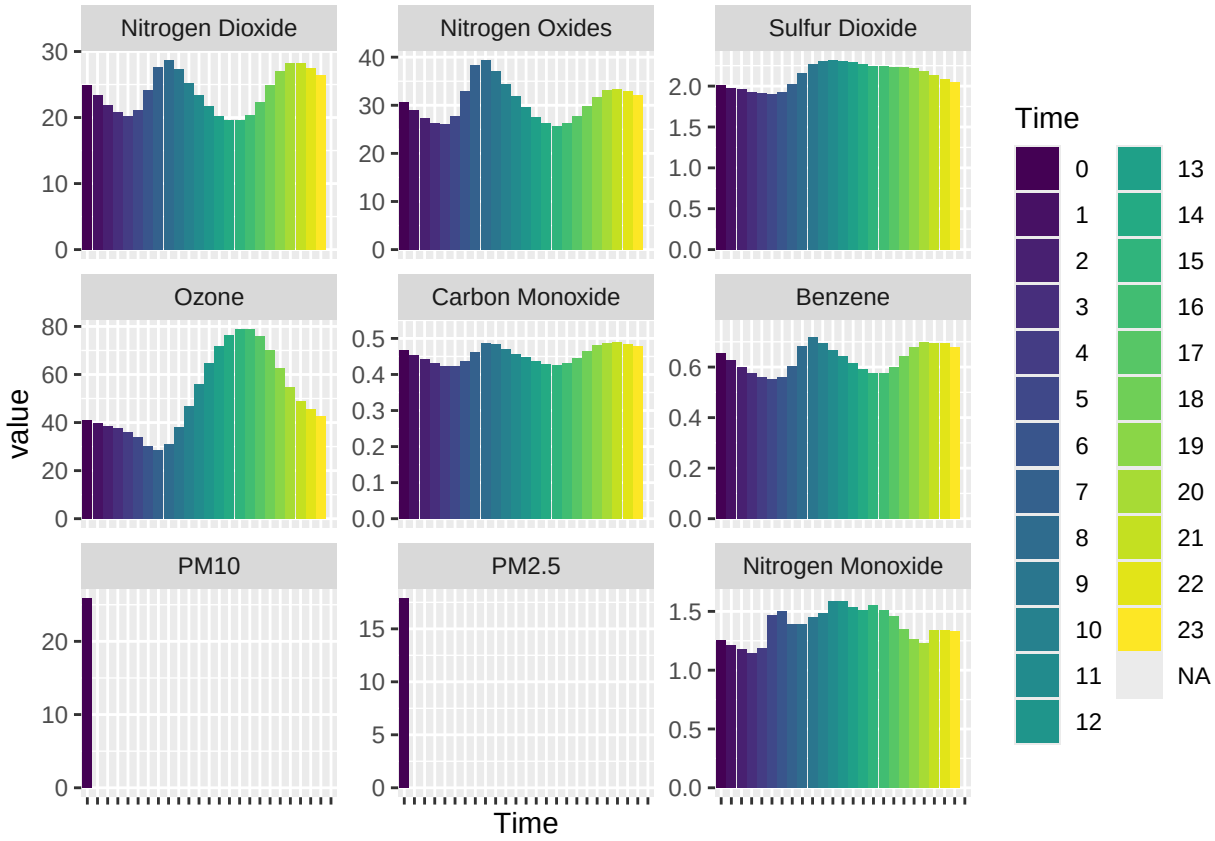


As we can see here some of the pollutants have a similar behavior, they have an higher concentration during the winter months, and a lower one during the summer. The exceptions are the Nitrogen Monoxide and the Ozone that has an opposite behavior respect the other, with an higher concentration during the summer.



From the graph we can see that the concentration of the pollutants it's not effected to much from the day of the week, it remains fairly constant for all days of the week. The only exceptions seems to be those related to nitrogen and PM2.5/10, with values lower during the weekends.

One explanation for the lower levels at weekends could be less traffic caused by people entering and leaving work.



First of all in the graph of the PM10 and PM2.5 there are only the the value related the midnight because, like stated from the source, their sensor communicate only the mean in the entire day, so it's correct to see the graph like that and unfortunately it's not possible to see how their level varies trough the day.

For the other pollutants it's possible to notice how in general for all of them, the concentration during the night is lower, and for some of them, like the Nitrogen Dioxide/Oxides, the Benzene and the Carbon Monoxide, we can see some high spikes in the 6-8 and in the 18-20. These bands coincide with the time of arrival and departure from work and probably are related with that and the heavier traffic in that hours. This may reconfirm what was seen in the previous example.

Studing the altitude

Now we will take in consideration the altitude of the sensor. For this first of all, to simplify the work, we are gonna add a new feature that approximate the altitude into a specific band. For example if a sensor is at between the 0 and 50 meter above the sea the value 50 will be assigned.

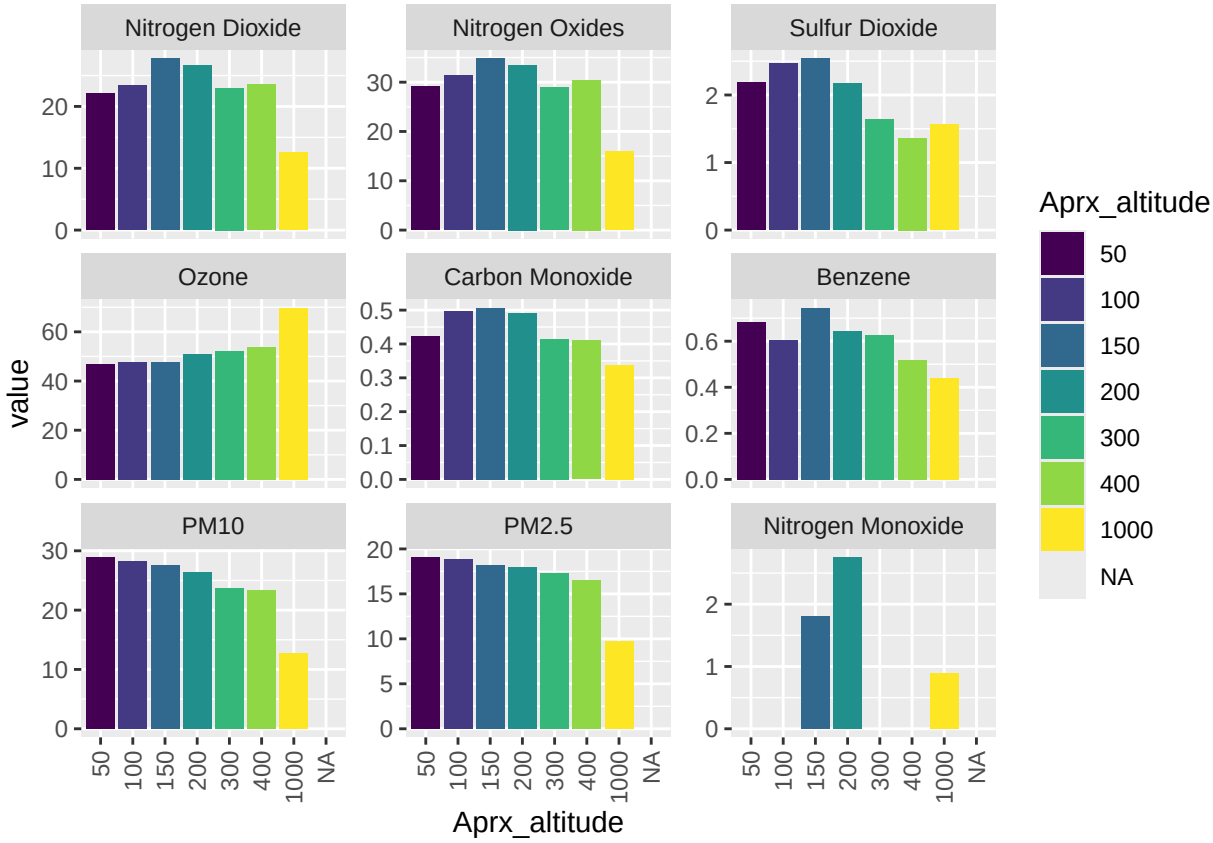
```
df$Aprx_altitude <- df$Altitude %>% {case_when(. > 0 & . < 50 ~ 50,
                                    . > 50 & . < 100 ~ 100,
                                    . > 100 & . < 150 ~ 150,
                                    . > 150 & . < 200 ~ 200,
                                    . > 200 & . < 300 ~ 300,
                                    . > 300 & . < 400 ~ 400,
                                    . > 400 ~ 1000)}
```

Then the graph of the averages is create like in the previous examples.

```
altitudes <- c(50,100,150,200,300,400,1000)
altitudes_average <- df %>%
```

```
  subset(select = c(Type,Value,Aprx_altitude)) %>%
  pivot_wider(names_from = Type, values_from = Value, values_fn = mean)
```

```
altitudes_average <- setDT(altitudes_average)
mtab = melt(altitudes_average, id.vars="Aprx_altitude")

mtab$Aprx_altitude = factor(mtab$Aprx_altitude, levels=c(altitudes))

p <- ggplot(data=mtab, aes(x=Aprx_altitude, y=value, fill=Aprx_altitude)) +
  geom_bar(stat="identity") +
  scale_fill_viridis_d() +
  facet_wrap(vars(variable),scales = "free_y") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))

p
```




```
ggsave("image/Aprx_altitude.png", plot=p, width=13, height=7, dpi=300)
```

From the graph it's possible to see how in general higher is the altitude lower is the concentration of the pollutant, with the only exception of the Ozone that has the opposite trend, being more present at the higher altitudes. This s probably due to the lower number of people and factories in that locations.

Pollutants level for provinces

Now we will see how the various provinces are ranked with respect to each pollutant. This time, we will not use bar graphs but will plot the intensity of the values directly on the map.

First of all, as before, the averages for each pollutant will be calculated.

```
provinces <- unique(df$Province)
provinces_averages <- df %>%
  subset(select = c(Type,Value,Province)) %>%
  pivot_wider(names_from = Type, values_from = Value, values_fn = mean)
```

After that we download the map on which we are going to report the results, to do this we will use the data provided by GADM, from which we will download the map of Italy at level 2, i.e. with regional and provincial borders, from which we will then extract the map relating only to the Lombardy region.

```
lombardy_region <- gadm(country = "Italy", level = 2, path = "./") %>%
  st_as_sf() %>%
  filter(NAME_1 == "Lombardia")
```

For the sake of simplicity, let us add a column with the full names of the provinces to the averages dataset. This will enable us later on to combine the map with the pollutant information by province.

```
NAME_2 <- c("Milano",
            "Bergamo",
            "Monza and Brianza",
            "Varese",
            "Como",
            "Sondrio",
            "Lecco",
            "Lodi",
            "Cremona",
            "Pavia",
            "Brescia",
            "Mantua")

provinces_averages$NAME_2 <- NAME_2
```

```
map_with_value <- left_join(lombardy_region, provinces_averages, by = "NAME_2")

plot_list <- list()
for (i in seq_along(pollutans)){
   c <- pollutans[i]
   p <- ggplot(map_with_value) +
     geom_sf(aes(fill = .data[[c]])) +
     scale_fill_gradient(low = "#a7b4e8",high = "#2e408c", name = c) +
     ggtitle(c) +
     theme(plot.title = element_text(hjust = 0.5))
   plot_list[[i]] <- p
}

#p <- grid.arrange(grobs=plot_list,ncol=3)
#ggsave("image/map_region.png",width=15, height=10, plot=p, dpi=300)
```

As can be seen from the figure, the distributions per pollutant are very different from each other: First of all, not all provinces have measurements for all pollutants, for example, for Nitrogen Monoxide only two, and another interesting thing to note is that in general, defining which province is the most polluted (overall) is not easy, some have high values for certain pollutants and low values for others, the only one that might stand out from the others is the province of Milan.

In general, however, we can see how most of the provinces, especially the southern ones, are affected by high values of PM10 and PM2.5 Suspended Particles, particularly in the province of Cremona. It is also

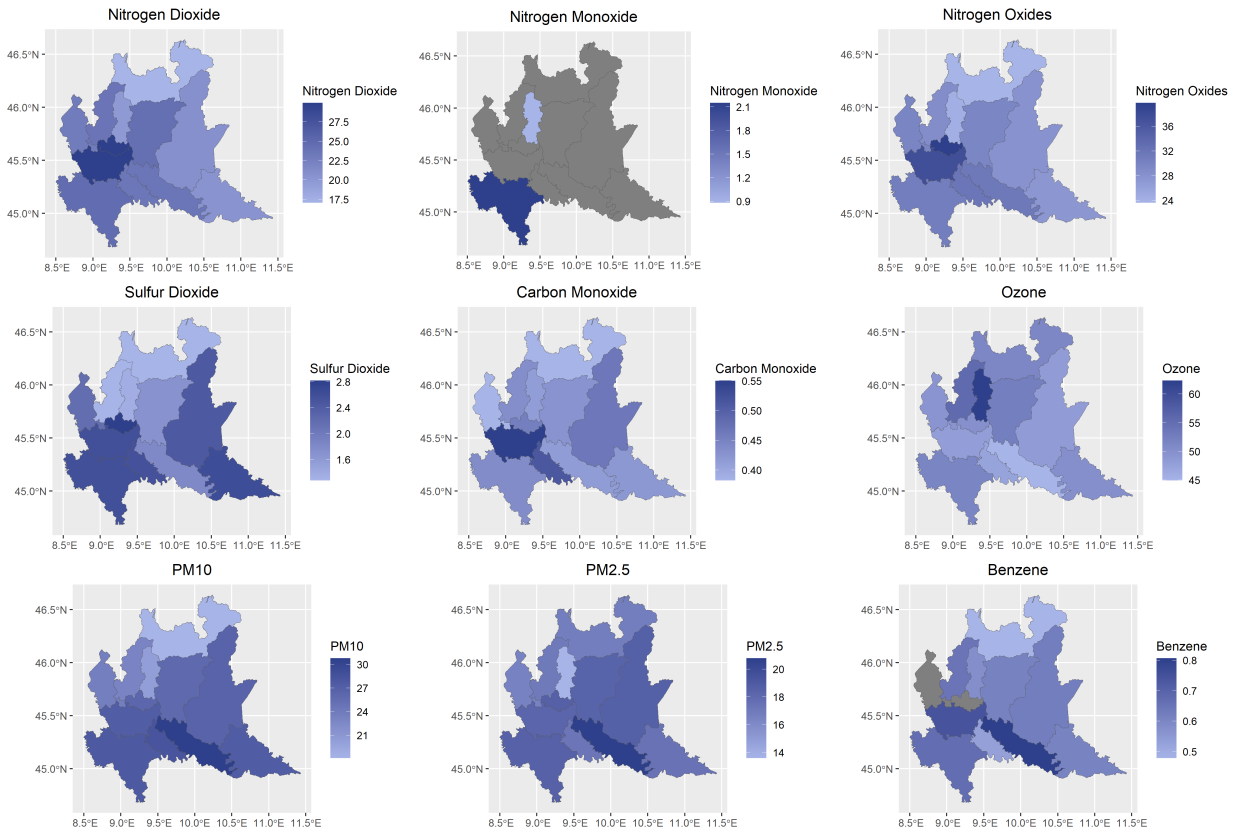


Figure 1: Provinces concentrations

interesting to see the Ozone situation, which is quite different from the others, with the province of Lecco being the most affected.

Finally, it can be seen that for Sulfur Dioxide the high values seem to be concentrated in a few provinces, namely Brescia, Mantua, Milan, Lodi, Pavia and Monza and Brianza. The remaining provinces seem to be far behind with much more restrained values.

A more detailed study at municipal level would have been interesting: for example, the Province of Brescia covers a very large area, and the values probably change a lot from area to area. Unfortunately, however, with such a study the map would have become much more fragmented and difficult to read, but above all, the main problem lies in the limited number of sensors, as many of the municipalities would be without surveys.

Pollutans and temperature correlation

From the previous graphs we have seen how the concentration of pollutants varies greatly as the time of year varies. Now we are going to take a closer look at these variations and temperature trends and see if the behavior is similar or if there may be correlations.

To do this, however, only temperatures in the province of Milan in 2018 will be considered. The analysis will be confined to this area because not all data were available in the region's portal.

https://dati.comune.milano.it/dataset/ds686-rilevazione-temperature-anno-2018

We extract from the dataset only data for the province of Milan in 2018.

```
milan_2018 <- df %>% filter(Date < "2019-01-01" & Province == "MI")
```

After that, for each pollutant we go on to generate the two graphs, one related to the levels of the detections, and one related to the temperatures, and place them next to each other to see the possible similar behaviors.

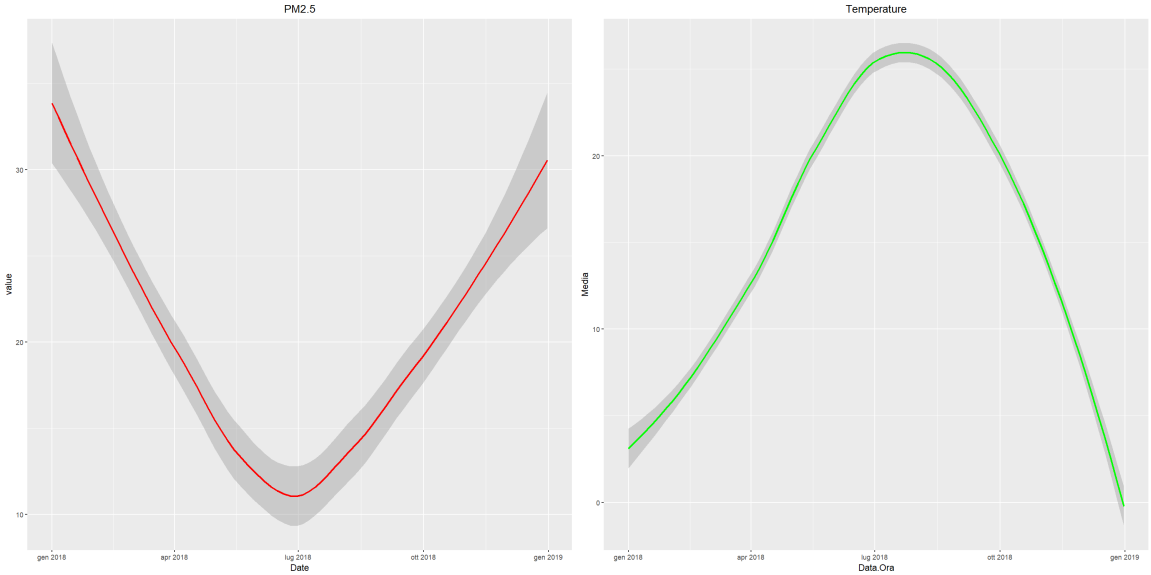
```
value_means <- function(df, pollutan){
  temp <- filter(df,df$Type == pollutan)
  temp <- temp %>%
    subset(select = c(Date,Value)) %>% # subset only the two column of interest
    group_by(Date) %>% # group by the date
    summarise(value = mean(Value)) # calculate the mean of each day
  temp$Date <- as.Date(temp$Date)
return(temp)
}
```

```
for (c in pollutans){
  temp <- value_means(milan_2018,c)
  p_list <- list()
  p1 <- ggplot(temp) +
    aes(x=Date, y=value, title(c)) +
    geom_smooth(color = "red")+
    ggtitle(c) +
    theme(plot.title = element_text(hjust = 0.5))
  p_list[[1]] <- p1

  p2 <- ggplot(milan_temps) +
    aes(x=Data.Ora, y = Media) +
    geom_smooth(color="green") +
    ggtitle("Temperature") +
    theme(plot.title = element_text(hjust = 0.5))
  p_list[[2]] <- p2
  p <- grid.arrange(grobs=p_list,ncol=2)
  path <- paste("./image/milan_",gsub(" ","",c),".png",sep="")
```

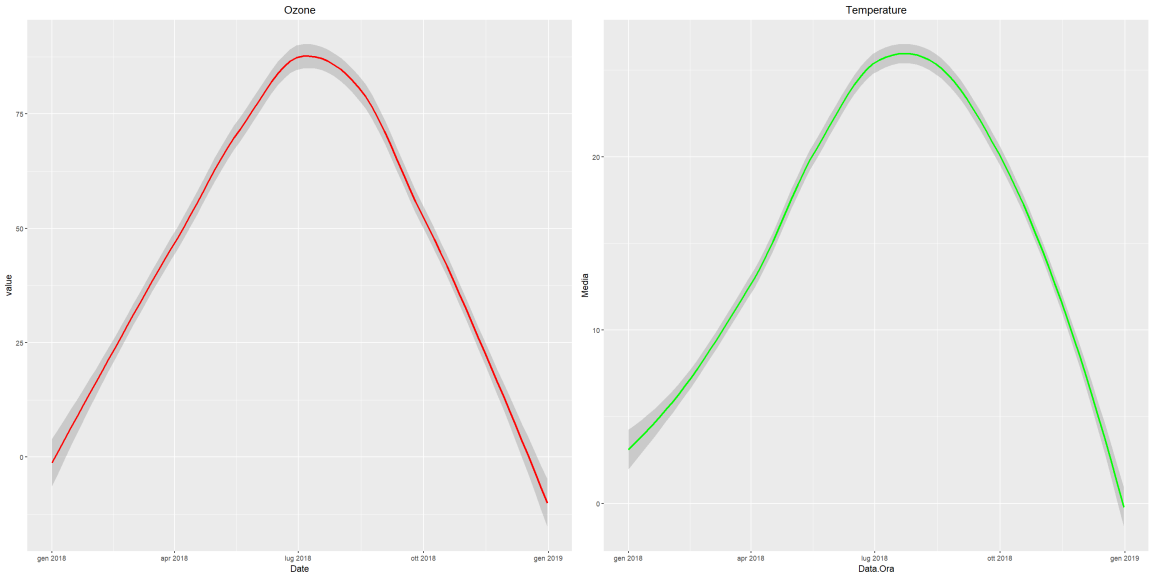
```
  ggsave(path, plot=p, width=20, height=10, dpi=150)
}
```

We will now look at some of the results as we basically saw from the previous graphs that there are two different behaviors.

One is where the concentration increases during colder periods. An example is the Suspended Particles PM2.5



The other behavior that has been observed is that of ozone, which has higher concentrations in summer periods.



As we can also see here as seen above most pollutants tend to decrease during periods when temperatures are higher, then opposite behavior during colder periods. One hypothesis then that could be made is therefore that one of the main polluting factors could be the heaters in homes and buildings in general. Another cause could be heavier traffic or industries and companies in higher productivity. This could explain the spike in winter periods and also the lack of decline during the pandemic period.

The only exception is, as seen above, Ozone, which exhibits the opposite behavior, with a curve similar to

that of temperature.

WHO guidelines and current situation

Instead, we are now going to compare the data obtained with the guidelines and to the interim targets provided by the WHO in the updated document in 2021.

Not all the pollutants previously considered will be present, but only those actually mentioned in the paper, in particular we will see: PM2.5, PM10, Carbon Monoxide, Nitrogen Dioxide and Sulfur Dioxide.

Each graph will show in black the pollutant concentration trend and a series of levels (Interim) that constitute the intermediate targets to be reached, the last line of which (in green) relates to the AQG level, the level to be reached for public health safety.

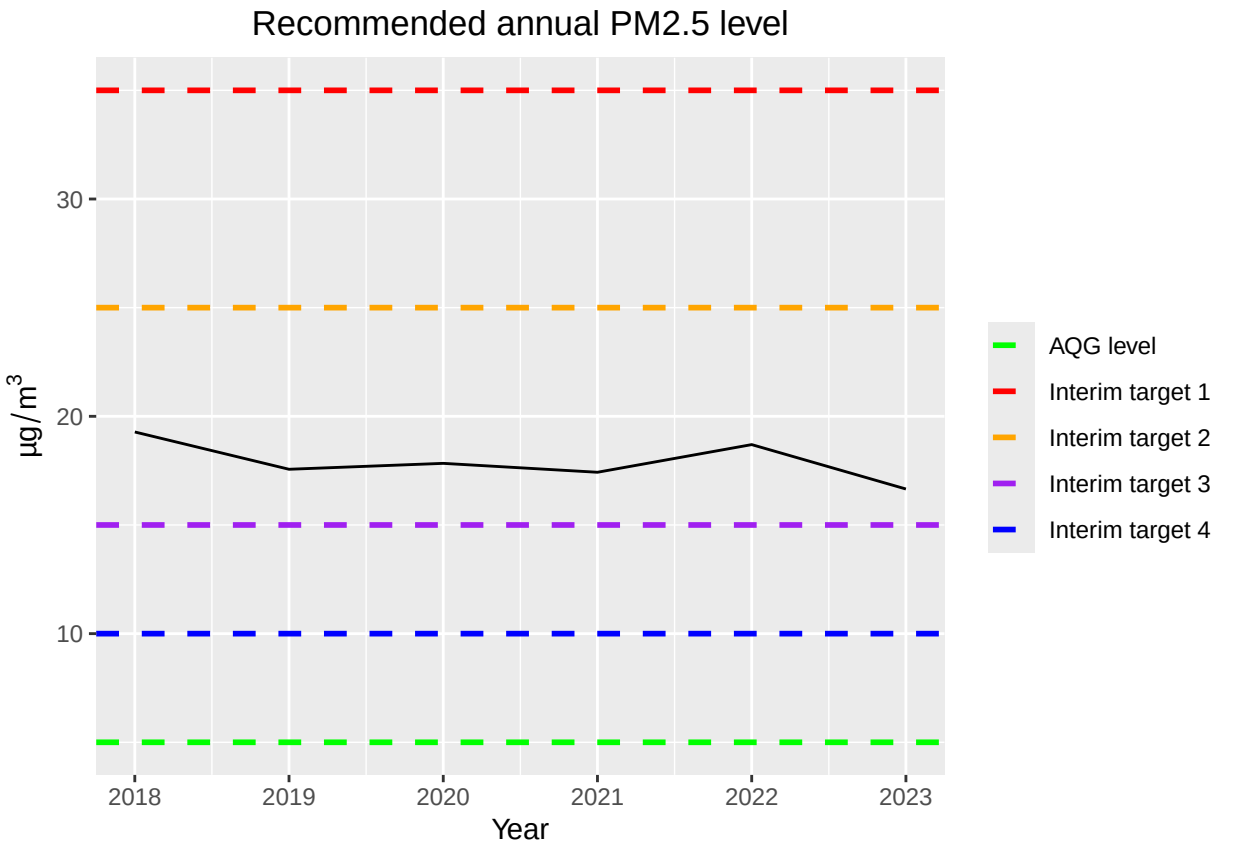
In general, a distinction is made between short-term and long-term exposure, so for each we will have two graphs, one relating to annual averages (long-term), and one relating to daily averages (short-term).

We will also analyse in more detail the evolution of the various pollutants over time, as taken individually it will be easier to notice abnormal behavior or improvement/worsening.

In the end, the code for each of these graphs is equivalent, so it will only be shown in the first case. Very simply, the averages calculated in previous studies will be reused, these values will then be shown together with the targets set by the WHO.

PM2.5

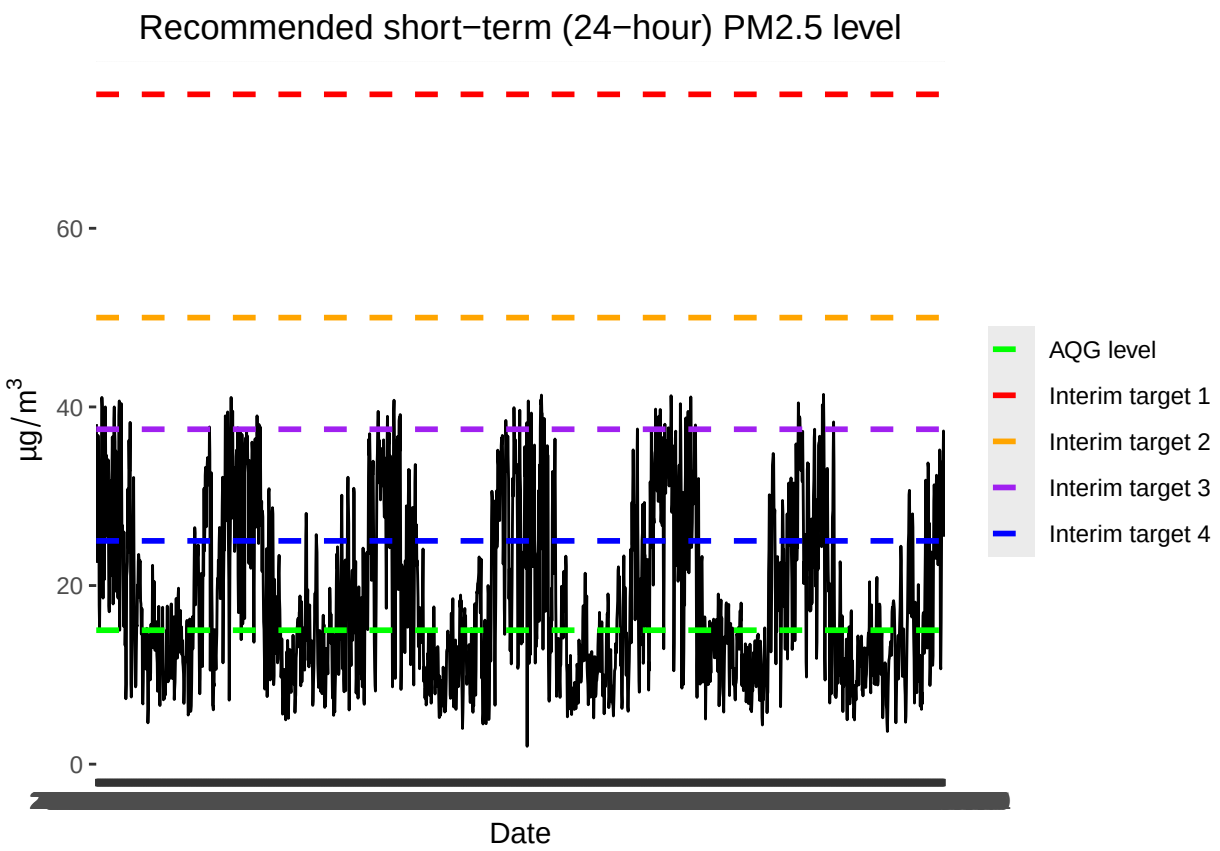
```
ggplot(annual_averages) +
  aes(x=Year, y=`PM2.5`, group = 1) +
  geom_line() +
  geom_hline(aes(yintercept=35, linetype="Interim target 1"), color = "red", size = 1) +
  geom_hline(aes(yintercept=25, linetype="Interim target 2"), color = "orange", size = 1) +
  geom_hline(aes(yintercept=15, linetype="Interim target 3"), color = "purple", size = 1) +
  geom_hline(aes(yintercept=10, linetype="Interim target 4"), color = "blue", size = 1) +
  geom_hline(aes(yintercept=5, linetype="AQG level"), color = "green", size = 1) +
  scale_linetype_manual(name ="", values = c('dashed','dashed','dashed','dashed','dashed')) +
  ggtitle("Recommended annual PM2.5 level") +
  ylab(expression(µg/m^{3})) +
  theme(plot.title = element_text(hjust = 0.5))
```



Regarding PM2.5 Suspended Particles as you can see the situation has not changed too much since 2018 but seems to be slightly decreasing. As far as the achieved targets are concerned, we are still between second and third, still far from the final target to be reached.

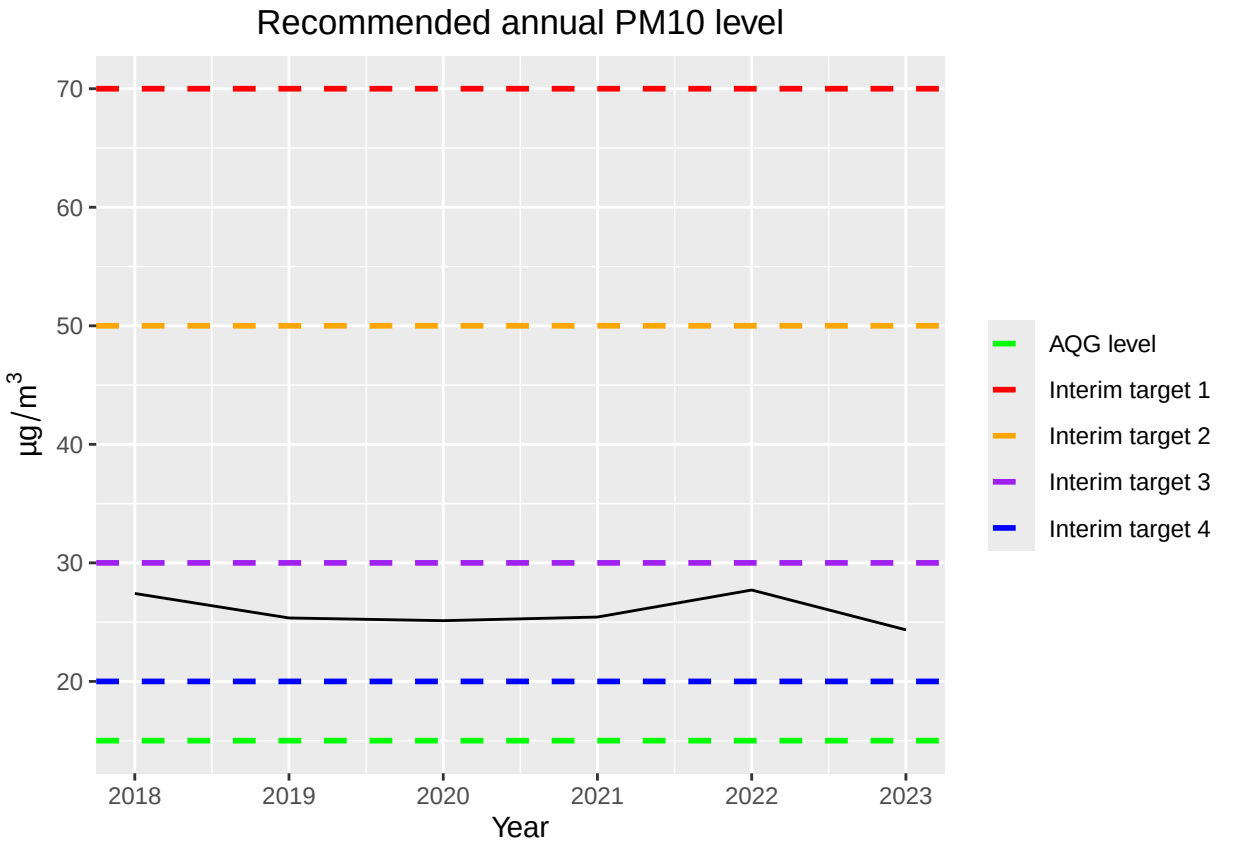
It is interesting to see how between 2019 and 2021 the levels dropped and remained more or less constant, then rose again in 2022, approaching the status of 2018, and finally falling again in 2023.

This behavior is probably not due to covid, the higher emissions in 2022 could be an effect of the general post-epidemic reopening, but this would not motivate the lower levels already in 2019, nor the final lowering in 2023.



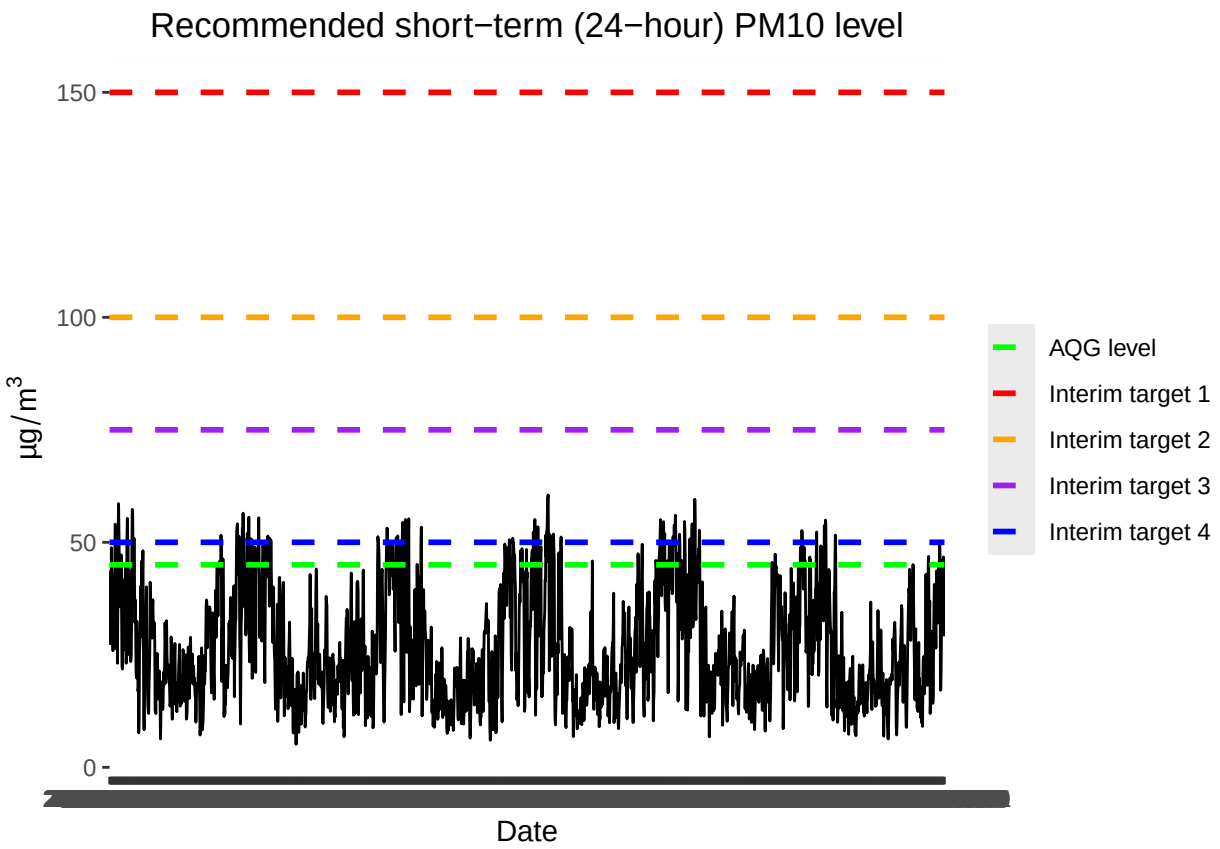
Looking at the daily averages, however, we can see that: during the summer months, readings also reach below the AQG level, during the winter months, on the other hand, during peaks, even the fourth target is exceeded.

PM10



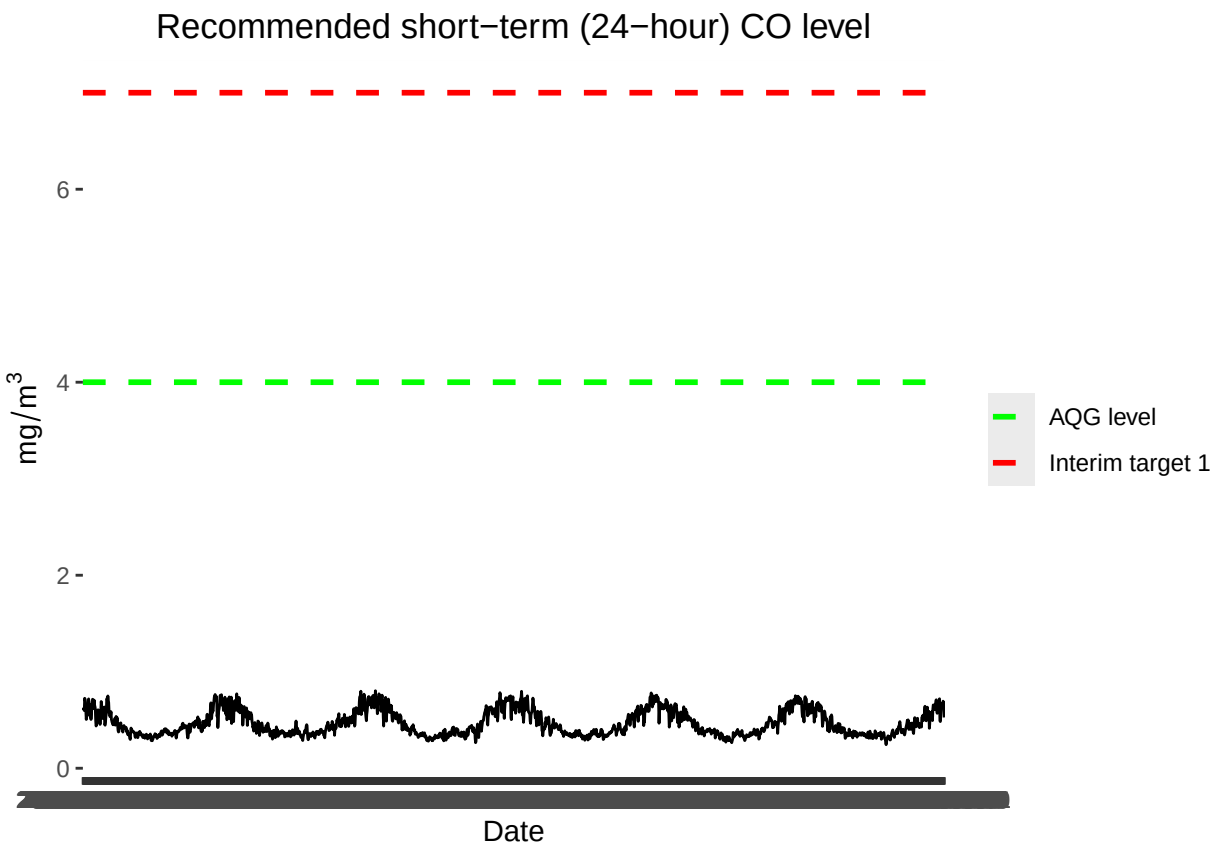
For PM10 Suspended Particles, for annual values, the situation is similar to the previous one: in each year we are between the third and fourth target, with a small peak during 2022.

Again, as before, one can see an anomalous behavior in 2022. One explanation for this peak could be the record drought recorded in that year. 2022 was the least rainy year since 1961, with the most marked anomalies being recorded in the north (-33% compared to the 1991-2020 climatological average) https://www.isprambiente.gov.it/it/archivio/notizie-e-novita-normative/notizie-ispra/2023/07/caldo-record-e-siccita-nel-2022. Low rainfall, therefore, could be one of the causes of the higher annual concentrations.



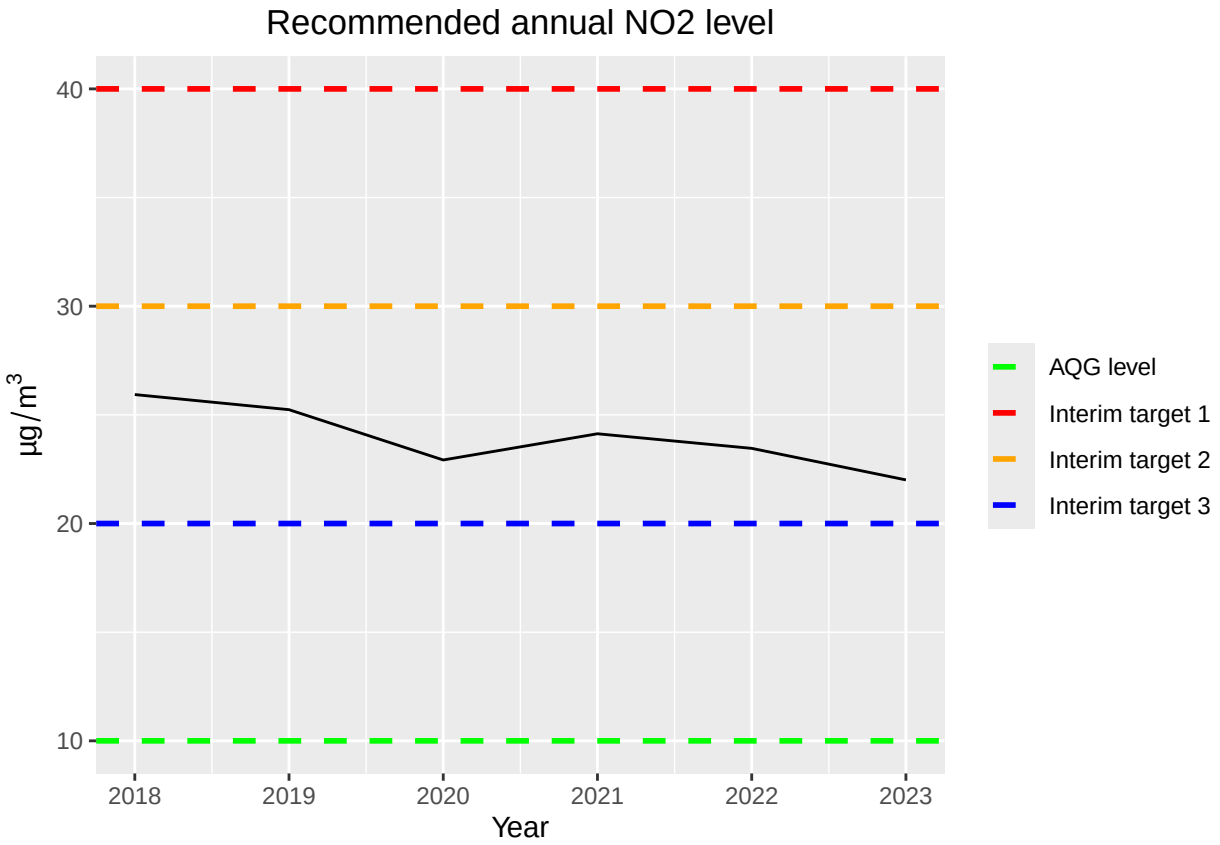
In contrast to PM2.5, the situation here seems much more positive: the averages seem to be below the AQG level much more consistently throughout the year, with the exception of the peak periods where the fourth target is slightly exceeded.

Carbon Monoxide



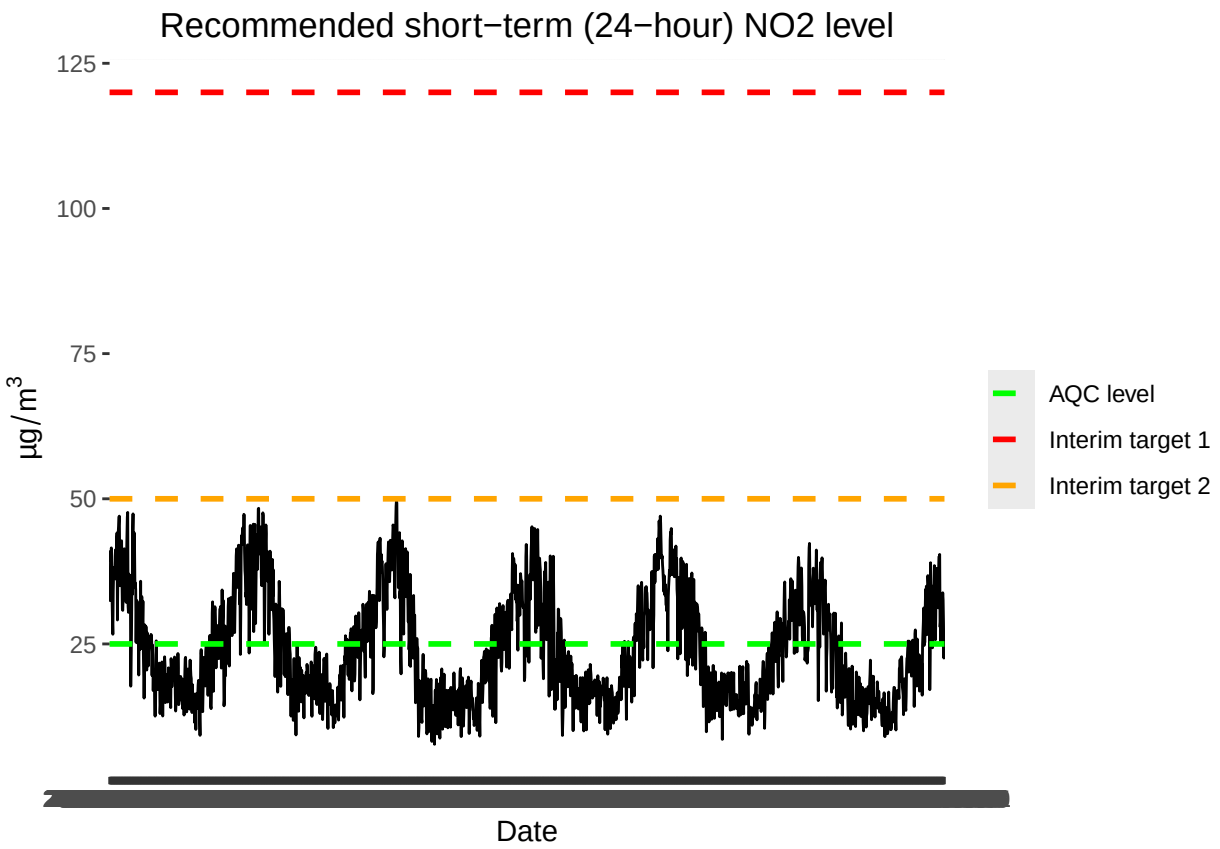
With regard to Carbon Monoxide from the guidelines we only have indications of short-term values, and there is only a target value and the AQG level, in any case as can be seen from the graph above the situation is well below the WHO recommended levels, with a fairly constant behaviour over the years.

Nitrogen Dioxide



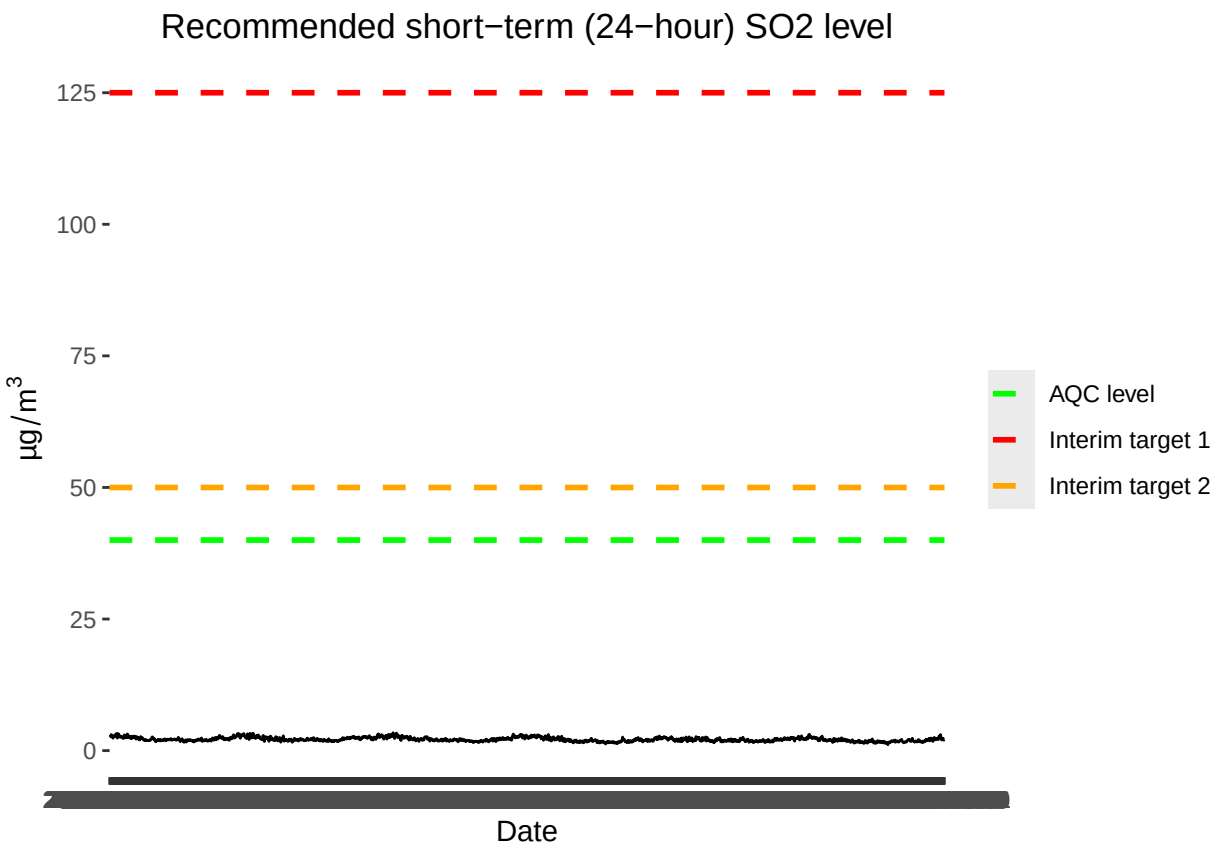
For Nitrogen Dioxide, on the other hand, the situation seems to be improving slightly: we are still between the second and third target, but the trend from 2018 seems to be decreasing slightly, getting closer and closer to the third target.

Here the anomalous behavior seen with PM2.5 and PM10 does not seem to recur, on the contrary a slight improvement over the previous year can be seen. On the other hand, it is interesting to see the negative peak recorded in 2020, this could be attributable to the pandemic, as the reduction compared to the previous year is considerable and is confined to the year 2020, the one most affected by the restrictions.



From the point of view of daily values, however, we are always below target, even during peak periods, and during the summer months below the AQG level. Even a slight improvement can be seen, a slight downward trend.

Sulfur Dioxide



As far as Sulfur Dioxide is concerned, again only daily values are given and as can be seen, the situation is always well below the target values and also well below the AQG level.

Conclusions

As we have seen, especially from the latest graphs, there is still a lot of work to be done with regard to air quality in the Lombardy region, especially with regard to particulate emissions, which, in some case, are still far from the levels recommended for public health.

It was interesting to see how, unexpectedly, the covid and pandemic months did not affect emission levels so markedly, except for Nitrogen Dioxide, despite periods of lockdown or limited mobility. One explanation for this could be the limited duration of these periods, we are talking in any case of a few months of total shutdown, in which some companies and activities remained active. Another explanation is that these emissions could also be linked to other factors, such as heating. In fact, the permanence of the population at home during those months could have contributed to a greater use of heaters and thus to a partial reduction in emissions due to the decrease in traffic. Proof of this is also the fact that the most severe lockdown periods, with the greatest restrictions, were the six winter months, also due to a greater transmissibility of the disease during those periods.

It would then have been interesting to compare the data with other data prior to 2018, extend this study, 5 years is a short time to see significant changes, barring significant events, with a larger time interval it would have been more interesting and more meaningful to see the evolution of emissions.