MASTER THESIS

Madrid Air Pollution Prediction

2º MASTER DATA-SCIENTIST
STREAMING 2021-2022

KSCHOOL

ALBERTO CORTINAS SANCHEZ

SUMMARY

1. INTRODUCTION

Pollution around the main urban agglomerations across the globe is one of the main problems that governments of all countries try to mitigate every year.

This situation was growing and growing during last centuries, reaching the current levels that are causing the death of more than 7 Mill of inhabitants per year (data from Paris climate conference [1]. WHO); a part of impacting in the right child development, increasing the risk of suffering from acute respiratory diseases, and developing chronic diseases.

In addition, the pollution has been identified as one of the main reasons of climate change.

Governments from the main countries are trying to take decisions to reduce it… but without a relevant success, just only agreeing some "cosmetics" actions that are not attacking the problem from the root, probably due to the complexity to balance those measurements together with the economic country development.

WHY

One of the tools available for the governments to mitigate and prevent is the monitoring of the main pollution parameters in the air and try to estimate them for the future to trigger preventive actions for the citizens.

WHERE

This study will be focused on Madrid, one the European cities with highest mortality in Europe due to the N02 index [2].

WHAT

The goal of this study is to create a prediction model, based on the historical Air Station and Weather data network from "Ayuntamiento de Madrid", with the objective to predict the pollution for the next 24 hours from the live data published for every Air Station.

This will be visualized in a web front-end that allow to the end user to monitor situation (almost on real-time) per AirStation and the related prediction for everyone for the next 24 hours.

---

[1] https://www.who.int/news/item/07-12-2015-health-savings-from-climate-mitigation-can-offset-costs-note-participants-at-cop21-paris-climate-conference

[2] https://www.elmundo.es/ciencia-y-salud/medio-ambiente/2021/11/11/618c5be5fdddff039e8b45e3.html

2. REPOSITORY AND ENVIRONMENT REQUIREMENTS

REPOSITORY

All the documentation and code are storage in a public GIT repository:
https://github.com/corti-acs/KS_TFM_MadridAirQuality

This repository is organized as follow:

- README.md. High level description of the repo, just for developers, or other data scientist interested to follow this study.

- requirements.txt. File with all libraries and their versions used in this study. This is used to allow any user replicate environment. (Instructions in specific section of this document)

- data. Folder to include the data used in the study. This is divided in 3 folders:
    - Raw: Original data and immutable data dump.
    - Interim. Intermediate data that has been already transformed
    - Processed. Data ready for modelling

- docs. This contains two folders:
    - General documentation of the study and Thesis Memo, documentation to give more context and relevant information about this study.
    - MemoThesis. Here is place where Master memory document is storage.

- models. This folder should storage the model used to predict pollution (pickle file), but due to size limitation in GIT this file is not available in the repository, this file can be found via dropbox in this link: https://www.dropbox.com/s/kt2k4qlmk69plww/poc_16_SARIMAmodel_301_311_12.pkl?dl=0 .

- notebooks. Here can be found all notebooks, with all the code needed in the study. Every notebook (or group of notebooks) contains the code of the different phases of a typical data scientist project. Data mining, Data Cleaning, Data transformation, Exploratory, Modelling, Prediction and Visualization.

- webapp This part of the repository contains charts and other visualizations like maps that will be used in the Webapp. Together with .py files will be needed to launch it. (Instructions are available in the addendum part of this memo)

ENVIRONMENT REQUIREMENTS.

The language used in this study is Python3.8, using Jupyter Notebook and multiple libraries. The libraries and versions used are documented in “requirements.txt”, available in GIT repository

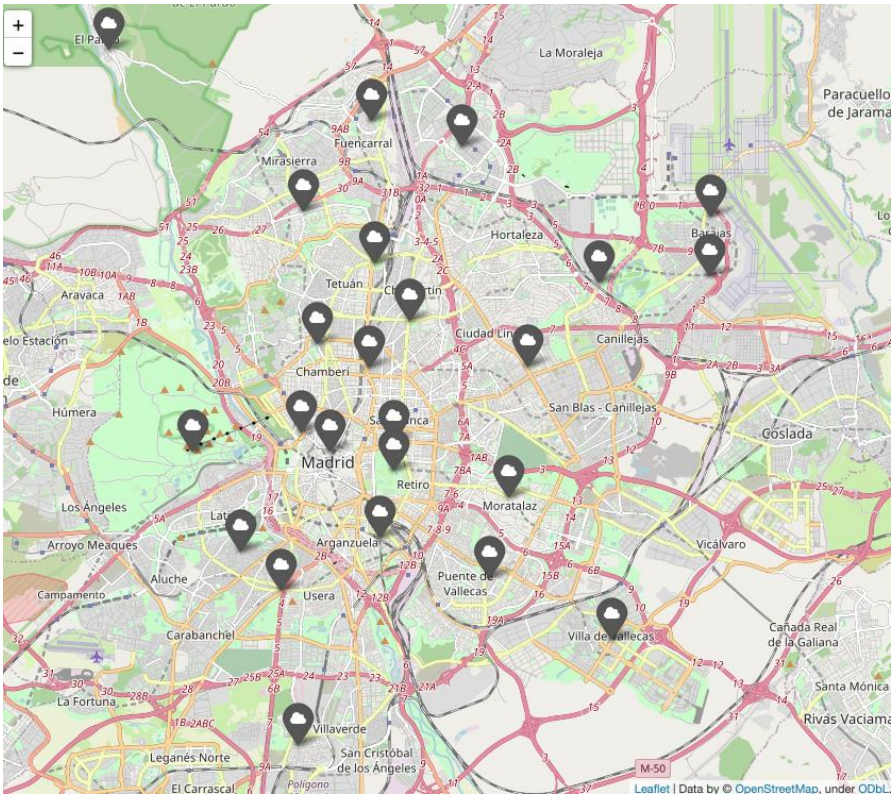
To replicate environment only it is needed to execute “!pip install -r requirements.txt” and automatically all libraries and versions used will be installed and will allow to replicate environment to execute the code done during the study.

```
Package                           Version
--------------------------------- -------------------
altair                            4.1.0
folium                            0.12.1
matplotlib                        3.3.2
numpy                             1.21.4
pandas                            1.1.3
pathtools                         0.1.2
pickleshare                       0.7.5
pmdarima                          1.8.4
requests                          2.24.0
scikit-learn                      0.23.2
scipy                             1.5.2
seaborn                           0.11.0
statsmodels                       0.13.1
streamlit                         1.2.0
```

3. DATASETS

All historical data used for analysis and modelling and live data for prediction used in this study is coming from the Open Data Portal from “Ayuntamiento de Madrid” [3]:

Ayuntamiento de Madrid has a network of 24 AirStations gathering till 8 different parameters to monitor pollution, although this study will focus only on the NO2 values



[3] https://datos.madrid.es/portal/site/egob

For this study 3 groups of data will be used.

- Historical AirQuality measurements. Ayuntamiento de Madrid has a daily (and hourly) dataset from 2001 to 2021 with different parameters measured. This information is published in different yearly zipfiles, containing everyone a monthly csv with all data. Data is published in txt, xml and csv.

  <u>Data used: csv files with data from 2019(January) to 2021 (September)</u>

  This data has the following format:

PROVINCIA	MUNICIPIO	ESTACION	MAGNITUD	PUNTO DE MUESTREO	AÑO	MES	DIA	Hxx (dato horario)	Vxx (código de validacion)
28	079	004	01	28079004_1_38	17	07	01	00005	V

  PROVINCIA. Spanish province where the study is located. 28 belongs to Madrid

  MUNICIPIO. Municipality where the study is located. 079 belongs to Madrid city

  ESTACION. Measuring station. In Madrid there are 24 stations across all the districts.

  MAGNITUD. Every station measures different parameters related to air quality. **This study will only need Magnitud related to Nitrogen Dioxide, NO2 that is represented as Magnitud = "08".** This is expressed in µg/m^3

  PUNTO DE MUESTREO. This is a code including the station id (including provincia, municipio y estacion), and adding the "magnitud" and "tecnica de muestreo".

  Hxx (dato horario). Data value per hour.

  Vxx (Código de validacion). Flag to identify the valid measurements, filtering out others that may have some issue during the measurement process. So, **the only valid values have a "V" in this field.**

  H01/VO1, H02/V02. There are 48 extra columns related to hour (H01 = 1:00 am, H02 = 2:00 am..) and validation (V01 = data flag related to the data at 1:00 am...). NOTE. For practical reason in the analysis the representation of the hours has been modified 1 second, just to allow reference the H24 to the same date. This is just matter of data representation not altering any result in the analysis.

- Historical Weather parameters measurements. Ayuntamiento de Madrid has another station network to monitor the most relevant weather parameters. Some of these stations are located together with the AirStations, but this not happening for all of them. Data Structure follow the same than AirStation.

PROVINCIA	MUNICIPIO	ESTACION	MAGNITUD	PUNTO DE MUESTREO	AÑO	MES	DIA	Hxx (dato horario)	Vxx (código de validacion)
28	079	004	01	28079004_1_38	17	07	01	00005	V

  PROVINCIA. Spanish province where the study is located. 28 belongs to Madrid

  MUNICIPIO. Municipality where the study is located. 079 belongs to Madrid city

ESTACION. Measuring station. In Madrid there are 26 stations across all the districts. Some of them are measuring also AirQuality.

MAGNITUD. Every station measures different parameters related to air quality. Below is the list of Magnitud and metric measured:

80 -> Ultraviolet radiation (MW/m2)
81 -> Wind speed (m/s)
82 -> Wind direction
83 -> Temperature (ºC)
86 -> Humidity (%)
87 -> Barometric pressure (mb)
88 -> Solar radiation (W/m2)
89 -> Water precipitation (l/m2)

PUNTO DE MUESTREO. This is a code including the station id (including provincia, municipio y estacion), and adding the "magnitud" and "tecnica de muestreo".

Hxx (dato horario). Data value per hour.

Vxx (Código de validacion). Flag to identify the valid measurements, filtering out others that may have some issue during the measurement process. So, the only valid values have a "V" in this field.

H01/VO1, H02/V02. There are 48 extra columns related to hour (H01 = 1:00 am, H02 = 2:00 am..) and validation (V01 = data flag related to the data at 1:00 am...). NOTE. For practical reason in the analysis the representation of the hours has been modified (reduced) 1 second, just to allow reference the H24 to the same date. This is just matter of data representation not altering any result in the analysis.

- Live AirQuality measurements. This is used to predict based on real-time measurements. Ayuntamiento de Madrid in its web[4] about AirQuality, provide measurements for all Air stations updated in real time, being the data updated every hour. This data update will happen between the minutes 20 and 30 of every hour. The data structure is the same than historical data described before.

  Disclaimer of live data: This data is live data automatically measured by the AirStations without any quality check, so data is pending to be reviewed and validated.

[4] https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=41e01e007c9db410VgnVCM2000000c205a0aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextfmt=default

4. METHODOLOGY

[5]

In all Data scientist project life cycle, there are common steps that needs to be followed in with efforts and dedication.

DATA MINING

(1a_Extract_Concat_AirQ_measurements.ipynb)
(1b_Extract_Concat_Weather_measurements.ipynb)
(2_Geo influence weather station per air quality station.ipynb)

This study is focusing in 2 type of data, AirQuality and Weather data, with both temporal (hourly) series it is expected to build a model good enough to predict pollution.
Study will use an hourly temporal series from 2019 to 2021 for AirQuality and Weather parameters.

AirQuality data:
RAW data is downloaded directly from the web, 1 zip file per year.
Every zip file contains 1 file per month, in 3 different formats (csv, txt, xml).
Automatically, all csv files are concatenated with the goal to create an only one dataset for AirQuality with the full temporal series

Weather data:
Same data structure and steps are followed for Weather data.

These reference files are storage for further use.

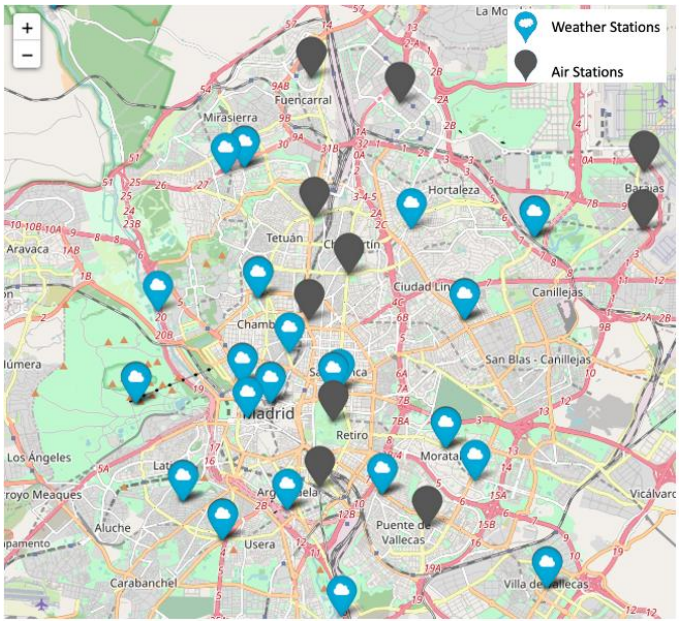
```
# Creation of weather parameter reference csv file
#####################################################

weather_df.to_csv("../data/interim/Weather/ref_weather.csv")
```

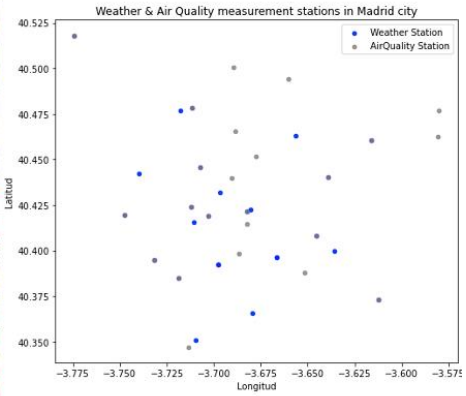
```
# Creation of reference AirQuality csv
#####################################

air_quality_df.to_csv("../data/interim/AirQuality/ref_air_quality.csv")
```

[5] https://medium.com/co-learning-lounge/complete-data-science-project-life-cycle-9eae6e4ed4c9

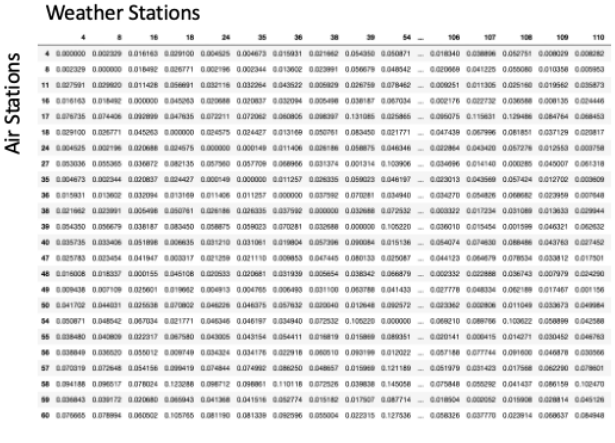
In this step there is other problem that should be tackled.
The intention is to link AirQuality data with Weather conditions and try to find a correlation somehow between some of them that can help to elaborate a prediction model based on them.

But the data from AirQuality and Weather conditions are not coming from the same stations, from the same places. There are 2 different station networks (26 stations measuring weather conditions and 24 stations measuring air quality) across Madrid Districts:





Using Euclidean distance[6], from SciPy library a distance matrix is created, to identify the Weather station closer to every Air Station.
Based on this matrix a reference table Air Station vs Weather Station is created.



scipy.spatial.distance.euclidean

```
scipy.spatial.distance.euclidean(u, v, w=None)
```

Computes the Euclidean distance between two 1-D arrays.

The Euclidean distance between 1-D arrays u and v, is defined as

$$\frac{\|u-v\|_2}{\left(\sum_i (w_i (u_i - v_i)^2)\right)^{1/2}}$$

AirStationCode	WeatherStationCode
4	4
8	8
11	38
16	16
17	103
18	18
24	24
27	109
35	35
36	36
38	38
39	39
40	56
47	112
48	16
49	110
50	107
54	54
55	107
56	56
57	39
58	58
59	59
60	58

[6] https://en.wikipedia.org/wiki/Euclidean_distance#:~:text=In%20mathematics%2C%20the%20Euclidean%20distance,being%20called%20the%20Pythagorean%20distance.

CLEANING AND TRANSFORMATION

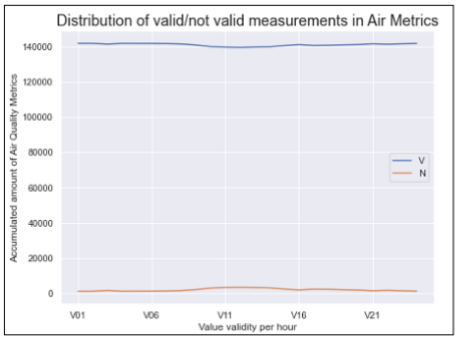
(3_Cleaning_Transformation_Exploratory)

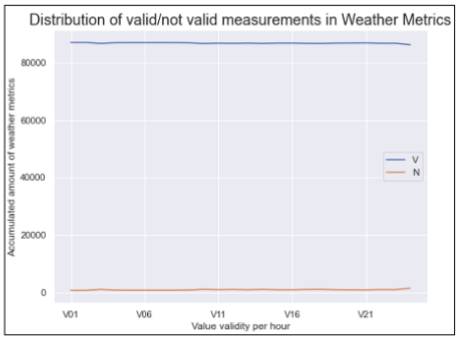
Using reference files created in the "Data mining" phase, now data needs to be transformed and cleaned to allow explore it and know more about it.

Different validations steps are done for both datasets, AirQuality and Weather

- Null data validation: First check is to verify that no NaNs in data. After verification we can verify that no column contains NaN value, in AirQuality neither Weather data.

- Valid Data: Provided data contains a "validation code" to allow flag measurements that due to some technical reasons are not valid, so first check is to identify if any measurement should be ignored. Valid Measurements are flagged with "V" and not Valid as "N"





V (=Valid)	N (=Not Valid)
3382952	48472
99%	1%

V (=Valid)	N (=Not Valid)
2082480	19944
99%	1%

Conclusion is that the Not Valid measurements is around 1% (in any of both datasets, Air and Weather data frame), so the non-valid values will be ignored due to the very low volume vs total, assuming that these values will not impact on the results of the study.

AirQuality dataset. Data Transformation

The scope of this study is only based on NO2 index, so, we should create a dataset only with this index. Which is identified in MAGNITUD field with value = 8.

The current data structure is not optimal for analysis, even not for modelling and predicting, so different transformations have been done:

- Creation of DATE field based on different time fields. This is done concatenating ANO, MES and DIA through "to_datetime" function.

- Wide to Long data frame transformation. Hourly measurements are storage in 24 columns, one per hour, this is not practical, so a new data transformation is required using "wide to long" approach, in this case, using "melt" function, This function is useful to massage a DataFrame into a format where one or more columns are identifier variables

(id_vars), while all other columns, considered measured variables (value_vars), are "unpivoted" to the row axis, leaving just two non-identifier columns, 'variable' and 'value'. https://pandas.pydata.org/docs/reference/api/pandas.melt.html#pandas.melt

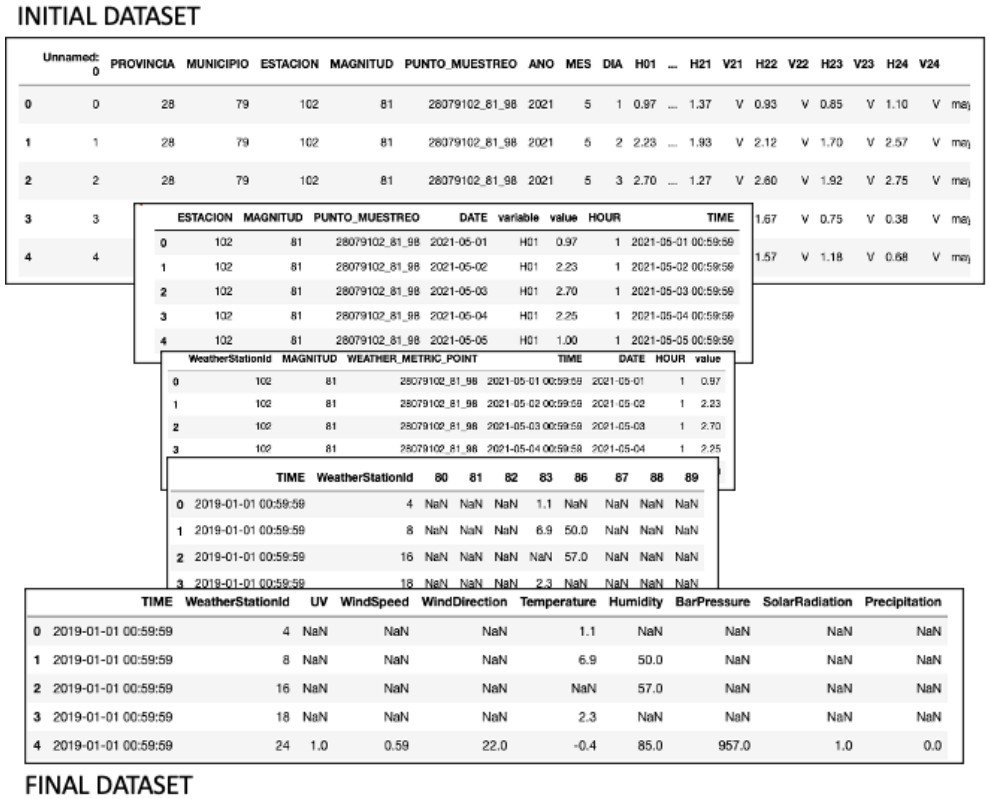
- Creation of Timestamp. This is the most relevant field, to be used as an index. Based on the hourly columns transformed using "melt" function, they are now transformed into a new HOUR field, that together DATE is getting TIME field (timestamp field). NOTE. HOUR is from H01 to H24, so for practical reason in the analysis the representation of the hours has been modified (reduced) 1 second, just to allow reference the H24 to the same date. This is just matter of data representation not altering any result in the analysis.

- Cleaning of non-useful fields. There are fields that are not relevant, so they are removed to do datasets more friendly/handy to work.

Weather dataset. Data Transformation

Data transformation is very similar than in the Air dataset. The first steps are the same, but this dataset has different parameters that we want to keep (precipitation, temperature, windspeed…), so it is needed to pivot the table to organize data. This is done moving different parameters (in MAGNITUD field) to different columns.

Overview of the dataset transformation:



But new Weather dataset is giving a situation where there a NaNs, so, this means that not all the Weather stations are reporting all weather parameters any time, a quick analysis is showing that parameters are far to be complete, except Temperature and Humidity, the rest of parameters are not registering more than 50% of the hourly data.

Columns	#Measurements with NaN	% NaN
TIME	0	0%
WeatherStationId	0	0%
UV	575712	98%
WindSpeed	356976	61%
WindDirection	357672	61%
Temperature	68160	12%
Humidity	65928	11%
BarPressure	402600	69%
SolarRadiation	401736	69%
Precipitation	355128	61%

This situation can have a big impact in the analysis (this will be evaluated in further steps of the analysis).

But thinking that weather data can help somehow to build a better prediction model, a strategy to fill NaNs have been defined, these NaN fields will be filled with the median from the rest of weather station at the same hour at the same day.

This strategy is feasible, except for UV index due to the big amount of NaN (85%); this is doing impossible to cover with any station all hours and days. So, as this parameter it looks like not to be relevant, and as we don't have other way to fill NaNs, so, this parameter has been excluded from the study.

Creation of data frame ready for modelling

Using the refence table created previously, it can be linked the Weather data to the Air data using the closer stations, creating a complete temporal series data frame with all NO2 index and weather parameter related to every Air Station.

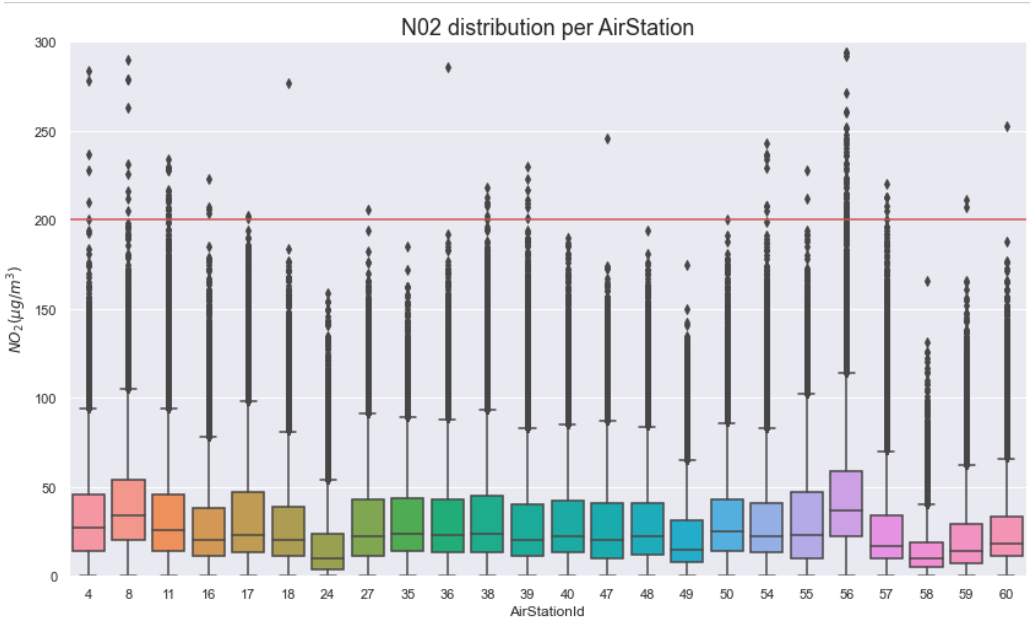
Creation of POC dataset

For the next phases of the study, we will focus the exploratory analysis and modelling in only one Air Station, considering that the results and conclusions of these phases are not going to be too different, as the area of study is in the same city.

The Air Station #16 located in Arturo Soria Street is the selected to be used. This was selected because it was looking for an average station, and this station is not at north, neither at south, not in the areas with more traffic, neither too far from the center.

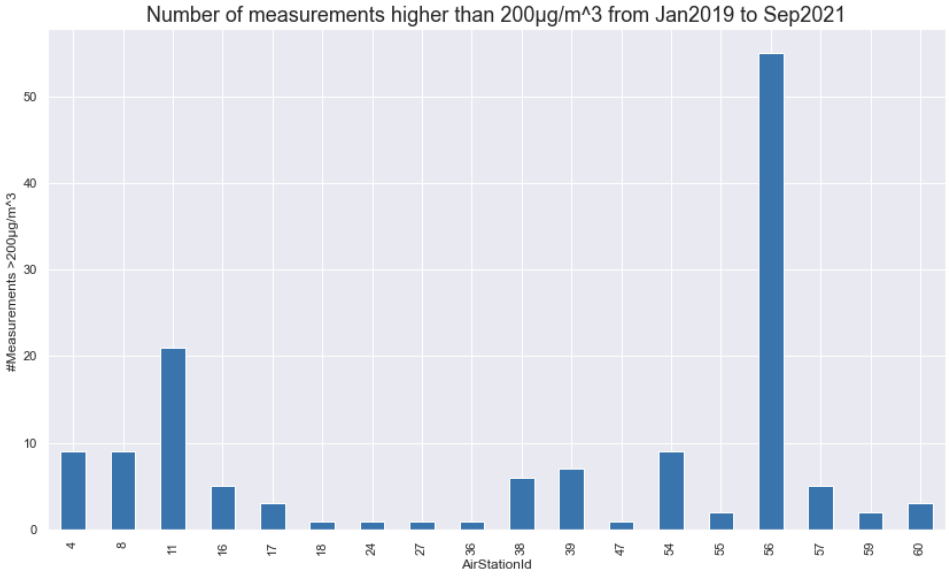
EXPLORATORY

N02 values behavior across all stations is very similar during the study period (from January 2019 to September 2021). Not having median values higher than 50 µg/m^3, and having limit periods of high pollution peaks (represented by values higher than 200 µg/m^3 , the red line).

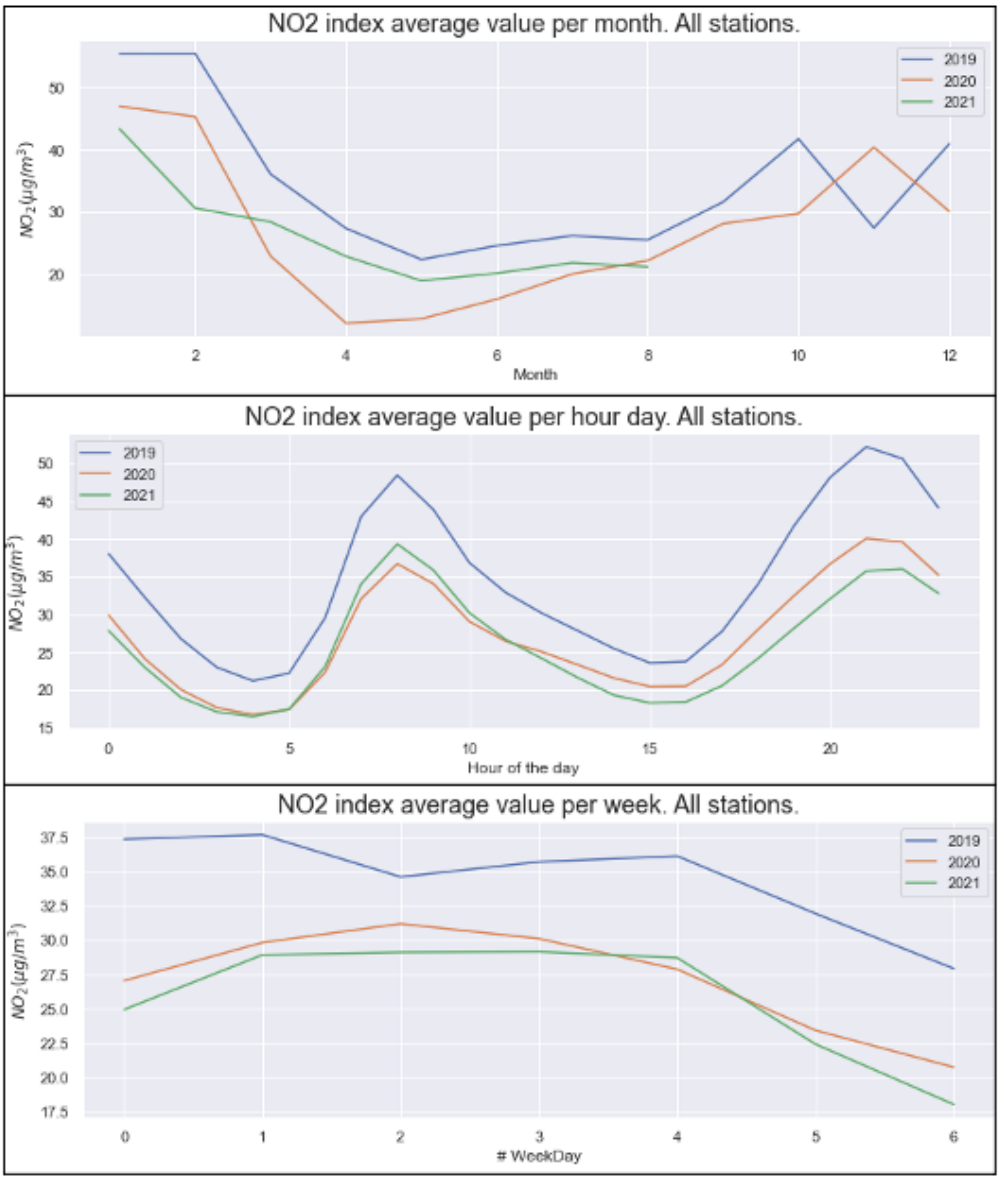


Being the stations with lower values the station 24 and 58, located in “La Casa Campo” and “El Pardo”

Although 18 from 24 Stations registered in any moment during study period some values higher than 200 µg/m^3, being the station 56 (located in Plaza Elíptica, the one where more often these high values are happening), followed by the station 11 (in Avenida Ramón y Cajal)



NO2 values, by default will be impacted by multiple parameters related to human habits: Peaks in moments and days with highest traffic congestion and linked to weather conditions which are more often in winter and autumn.



In the case of our POC Air Station (16) the pattern and the evolution are very similar to the rest of Air Stations.

- **Air Station 16 vs Rest of Stations.** Air Station 16 has similar values than other stations, so it can be used as POC for modeling, expecting that the accuracy of the model will work in similar way for any Air Station.

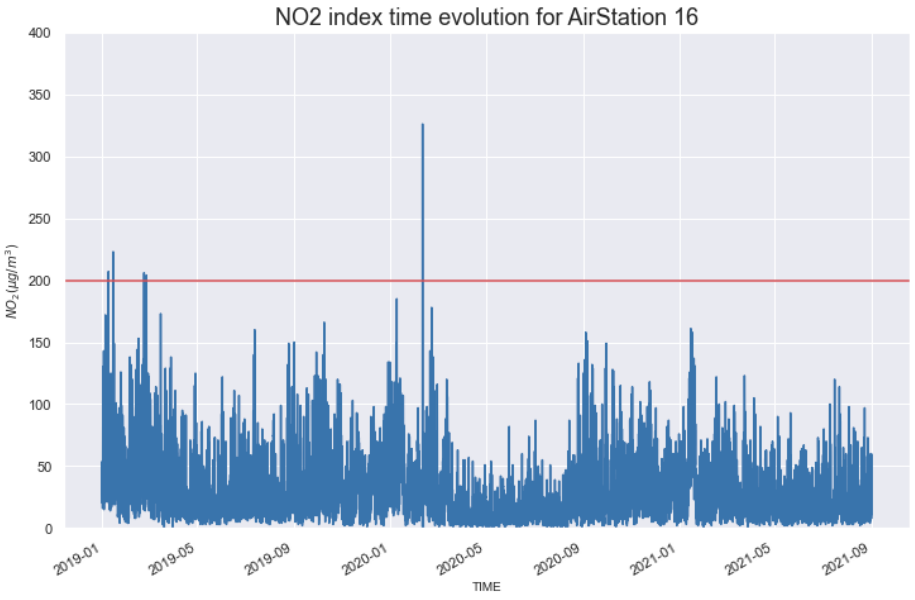
All Stations

	count	mean	std	min	25%	50%	75%	max
NO2_index	551928.0	29.714013	26.119170	0.0	11.00	21.00	41.00	1162.0
UV	106992.0	15.138188	28.215979	0.0	1.00	1.00	16.00	199.0
WindSpeed	725568.0	1.314200	0.910832	0.0	0.62	1.12	1.78	10.8
WindDirection	725568.0	142.049529	91.788965	0.0	55.00	139.00	225.00	360.0
Temperature	725568.0	15.751527	9.008342	-55.0	9.20	14.60	22.20	61.9
Humidity	725568.0	54.253492	23.923446	-25.0	35.00	53.00	74.00	100.0
BarPressure	725568.0	942.632170	35.844496	0.0	941.00	944.00	948.00	1137.0
SolarRadiation	725568.0	203.084538	289.803699	0.0	0.00	9.00	367.50	3789.0
Precipitation	725568.0	0.034504	0.335313	0.0	0.00	0.00	0.00	30.4

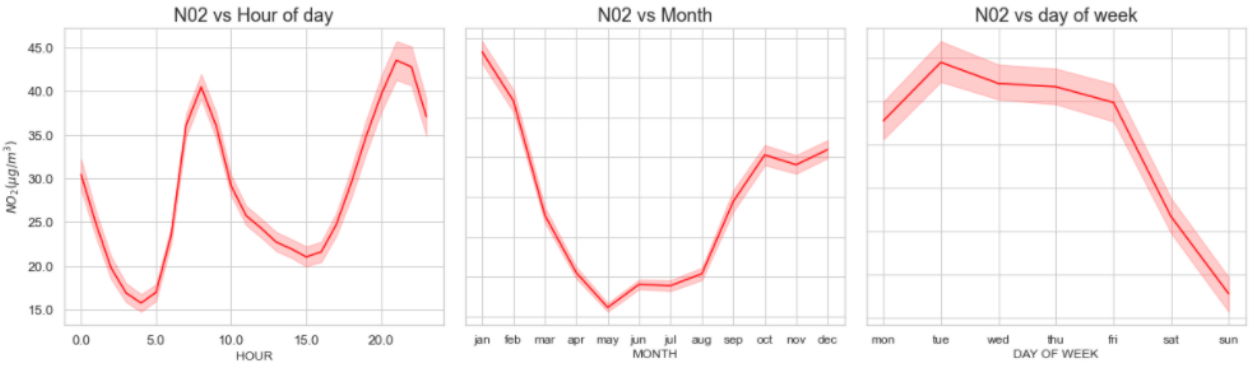
Station 16

	count	mean	std	min	25%	50%	75%	max
NO2_index	23376.0	28.315495	25.121162	0.0	11.00	20.00	38.0000	326.000
UV	3360.0	15.109524	28.067369	0.0	1.00	1.00	16.0000	160.000
WindSpeed	23376.0	1.302648	0.846464	0.0	0.62	1.12	1.7650	6.645
WindDirection	23376.0	140.083954	85.985391	0.0	57.50	136.00	222.0000	342.500
Temperature	23376.0	16.333415	8.653164	-8.1	9.70	15.20	22.7125	41.300
Humidity	23376.0	55.041068	22.883869	-25.0	36.00	54.00	74.0000	100.000
BarPressure	23376.0	943.796158	13.286819	0.0	941.00	944.00	947.0000	958.000
SolarRadiation	23376.0	204.843707	290.644003	0.0	0.00	10.00	372.5000	1038.000
Precipitation	23376.0	0.030409	0.271909	0.0	0.00	0.00	0.0000	11.300

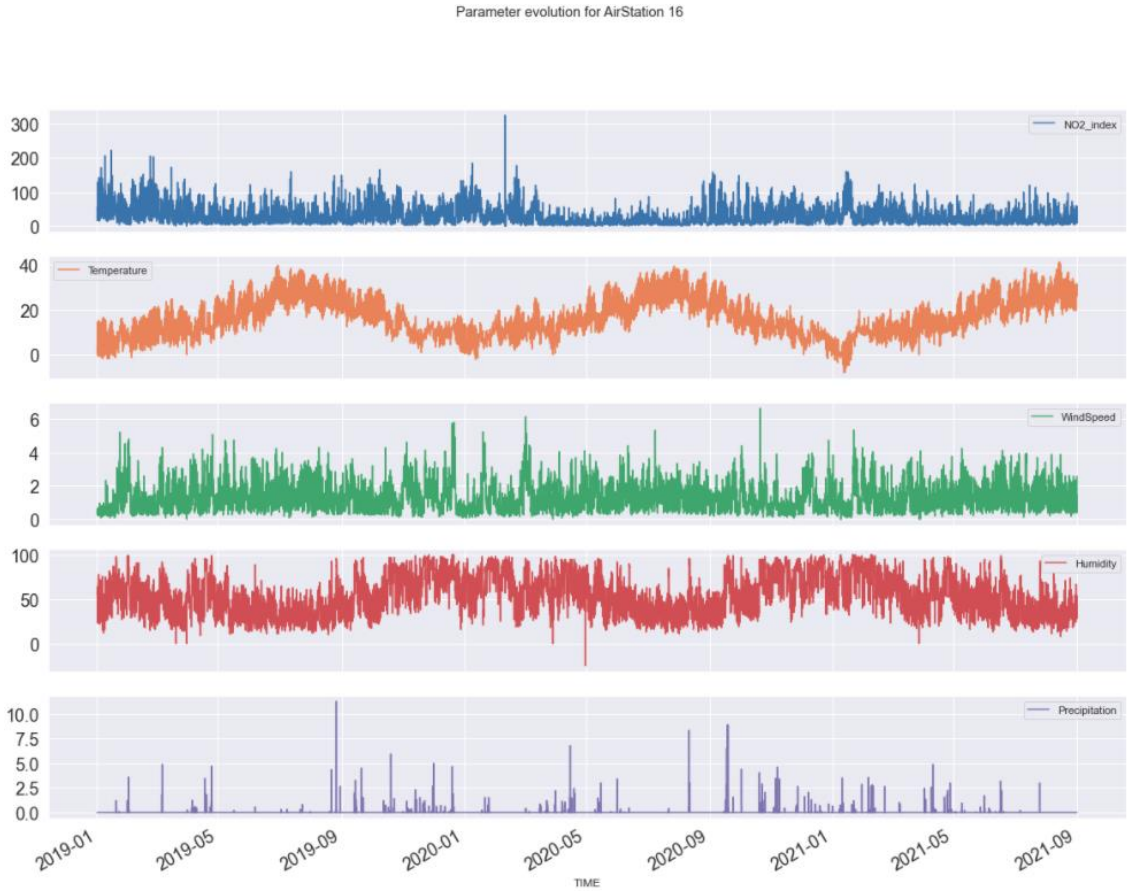
- **NO2 evolution.** N02 values don't use to reach higher values than 200 µg/m^3 (health limit). Only 5 measurements were above in 2 years and 8 months.



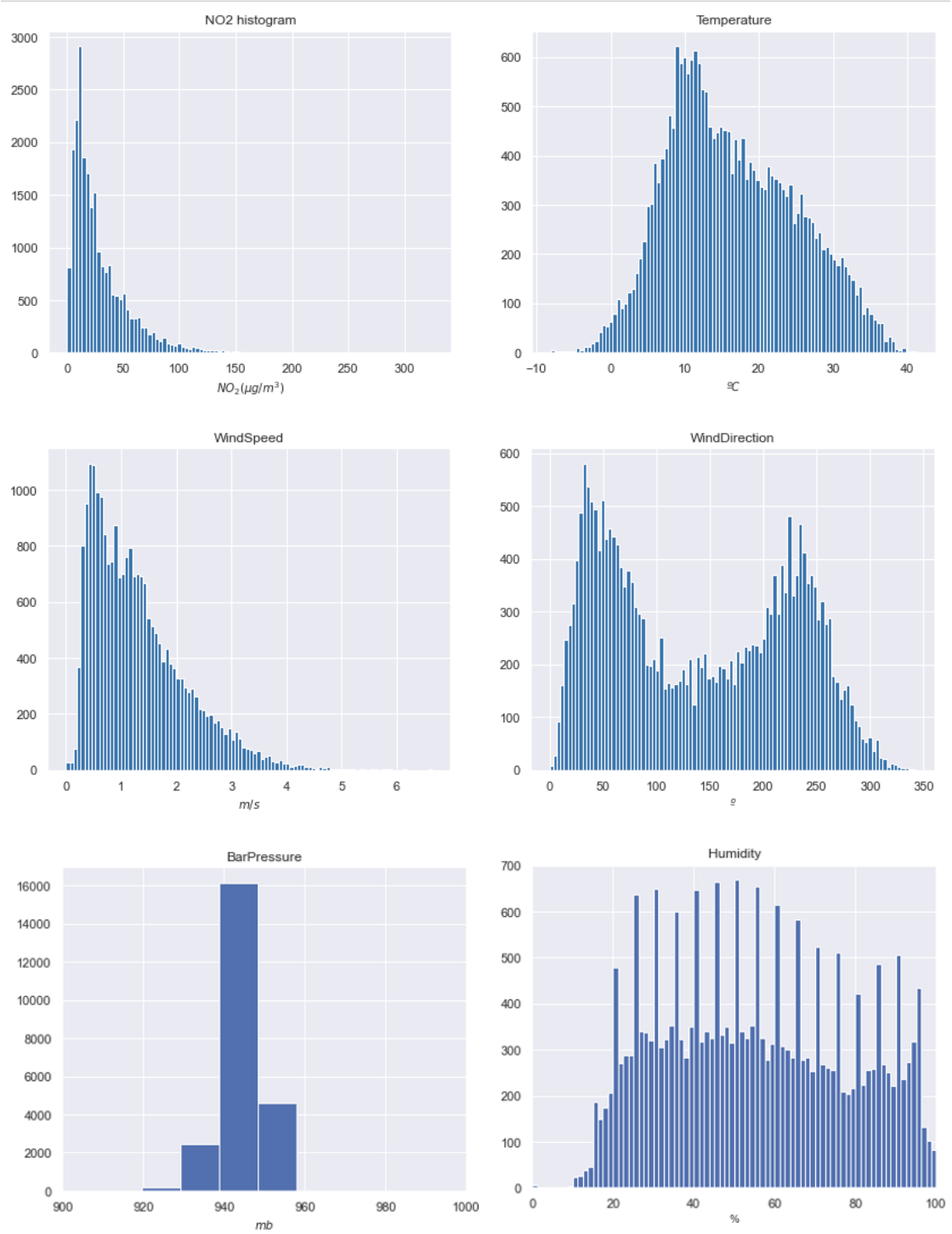
- **NO2 some patterns.** NO2 values follow different pattern depends on hour, day of week and month.



- **Parameter Evolution.** Like the NO2 peaks are short events, it is not possible to get any correlation in below charts.

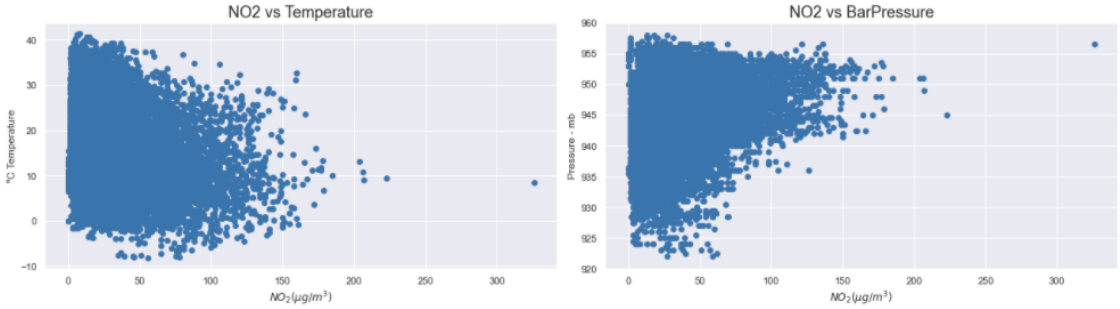


- **Histograms.** Histograms are giving an overview of the distribution of the different values for the most relevant parameters of the station.

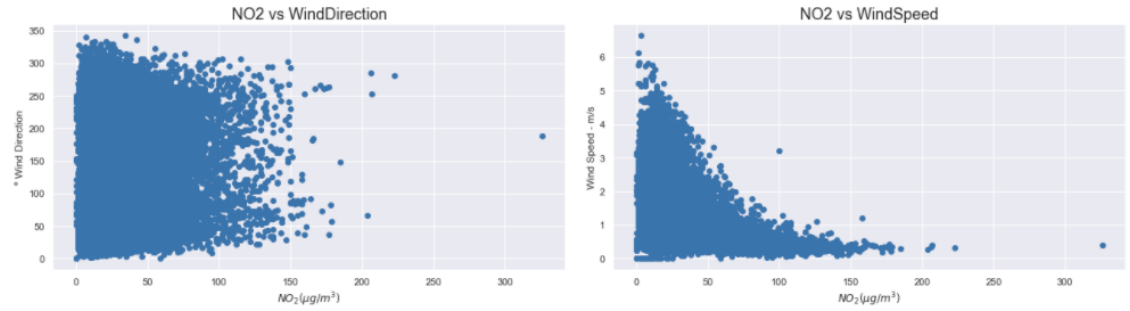


- **Correlation scatter plots.** But it is more relevant the information that different scatter plots are giving about the correlation of these weather parameters vs NO2 values.

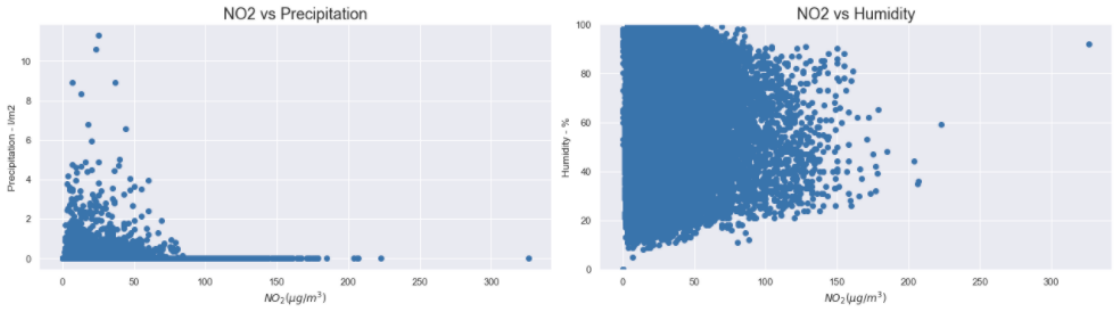
-Temperature between 8º and 15º give slightly higher N02 values
-Higher Bar Pressure values have more probability to give higher N02 values



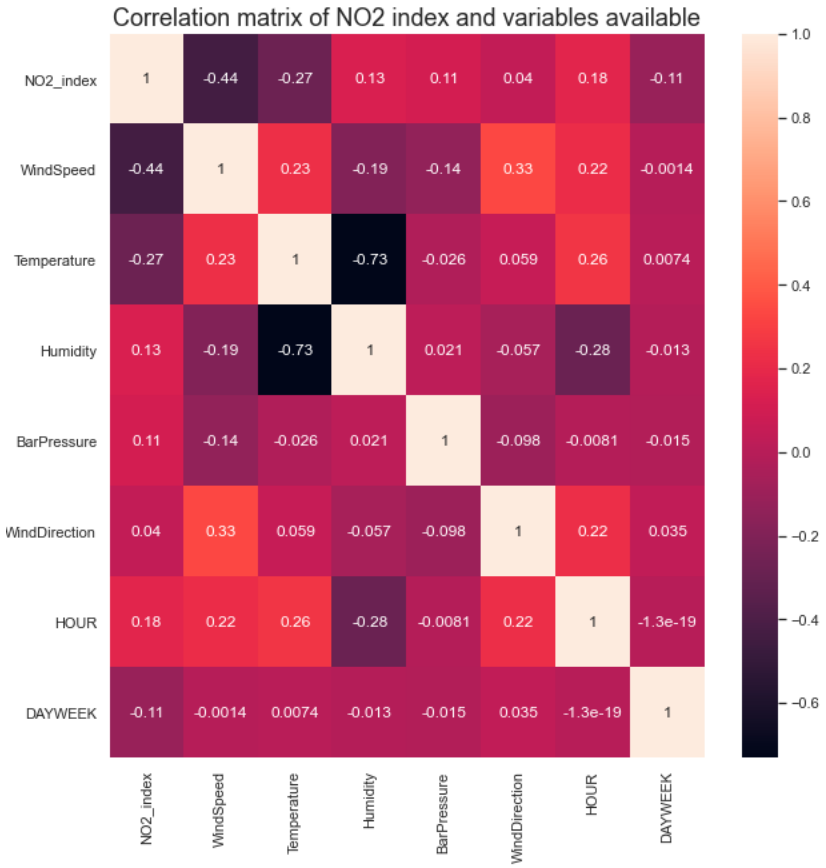
-Slow wind speed increase the values of NO2, although wind direction looks independent vs NO2 values.



- Higher values of NO2 appear more often with the absence of precipitations, although it is not impacted by medium humidity values.



- **Correlation.** To find correlation between the different variables, a correlation matrix is a useful methodology, in this case, Person's correlation coefficient [7] is used.



Although the previous scatter plots were showing certain correlation between NO2 and some variables (like Temperature or Wind Speed), this is not so big, having correlation coefficients of -0.44 and -0.26, what show not so big strong correlation. Something that we will have to consider during model creation.

[7] https://www.sciencedirect.com/topics/computer-science/pearson-correlation#:~:text=The%20Pearson%20correlation%20measures%20the,meaning%20a%20total%20positive%20correlation.

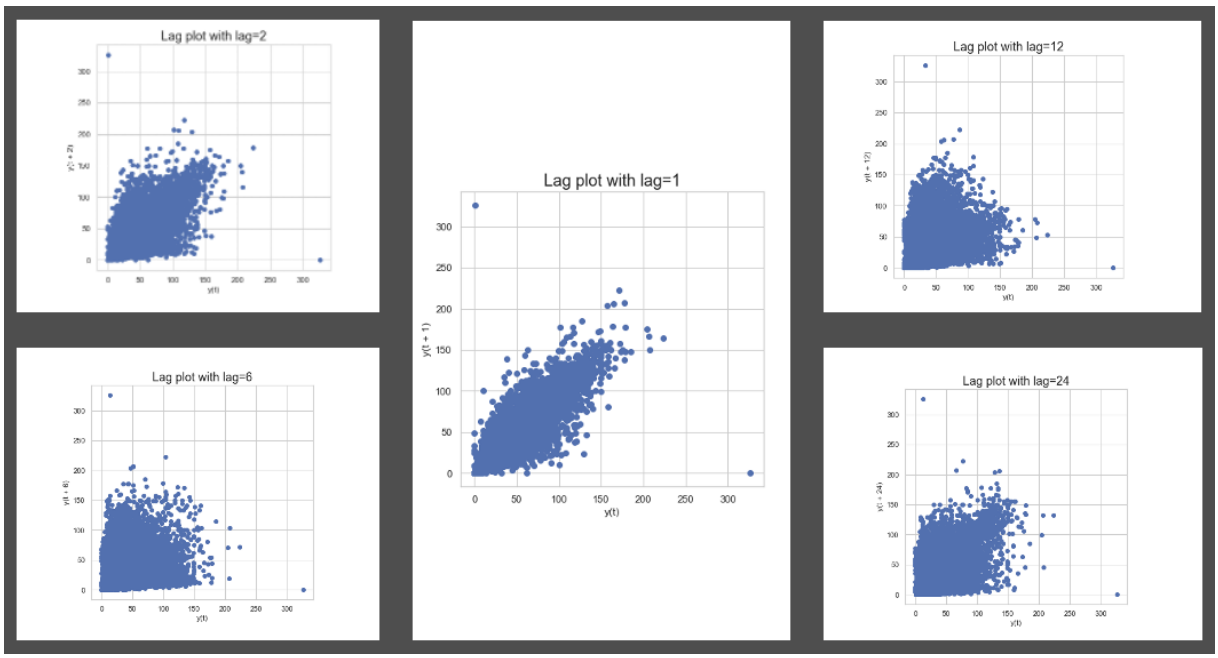
- LAG. In addition to the correlation with other parameters, other important aspect to consider for modeling is the lag factor between different measurements. A lag in a time-series data is how much one point is falling behind in time from another data point.

  In a time-series data, data points are marked over time at varying degrees of intervals.

  To analyze and find out if a time-series data follows any pattern, a lag plot can be employed.

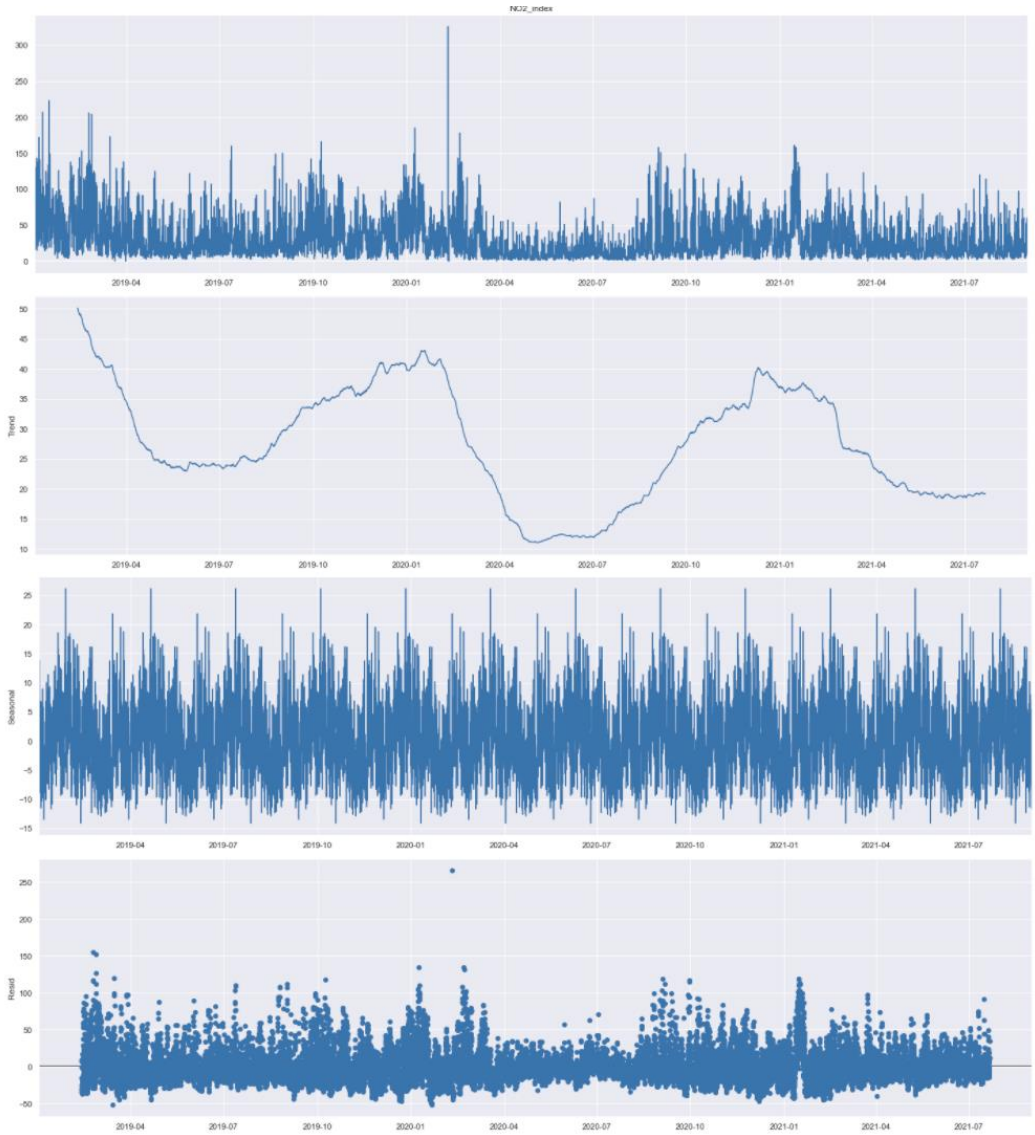
  A lag plot is drawn by representing the time series data in x-axis and the lag of the time series data point in y axis. For a data point, if the order of the lag is one, the lag is the previous data point. If the lag is two, the lag is data point before two data points in time.

  By drawing a lag plot, patterns like randomness, trends and seasonality can be searched for.



This time series is not randomly distributed, following a linear pattern, more linear as less lag has.

- **Seasonal decompose** [8]. This procedure decomposes a temporal series into their seasonal components, putting together:
    - Data
    - Trend. The increase or decreasing values in the series.
    - Seasonality. The repeating short-term cycle in the series.
    - Noise. Random variation of the series



Above charts are giving information about the NO2 time series value, showing a slightly negative trend, having the peaks a bit lower than previous.
Serie has a clear component of seasonality, but also have a relevant amount of noise, probably due to pollution peaks are happening randomly when different conditions push the NO2 levels.

This series parameters are indicating SARIMAX as one of the best to use for prediction.

[8] https://machinelearningmastery.com/decompose-time-series-data-trend-seasonality/

MODELLING & PREDICTION

(4_Modelling_SARIMA_poc16.ipynb)
(4_Modelling_EXOG_SARIMA_poc16.ipynb)
(5_LiveDataPrediction.ipynb)

ARIMA[9] is one of the most used models used for time series forecasting., ARIMA stands for Autoregressive Integrated MovingAverage, this model explains a given time series based on its own past values.

This temporal series has a visible seasonality distribution (see seasonal decompose chart in previous section), as standard ARIMA model doesn't support seasonality, SARIMA will be used (it is an ARIMA but adding seasonality parameters)
This seasonal differencing is just like regular differencing, but instead of subtracting successive terms, it subtracts the value from previous season.

ARIMA model includes three main parameters as p, d and q.

p: The order of the autoregressive model (the number of lag observations, or lag order).

d: The number of differences required to make the time series stationary, in other words, the number of times that the raw observations are differenced.

q: The order of the moving average model (size of moving average).

The SARIMA model builds upon the ARIMA model. It also includes the p, q, and d parameters, but also an extra set of parameters to account for time series seasonality.

P: The order of the seasonal autoregressive model.

Q: The order of the seasonal moving average model.

D: The number of seasonal differences applied to the time series.

Therefore we can denote the notation of the SARIMA model as SARIMA(p,d,q)(P,D,Q,s).

[9] https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/
https://towardsdatascience.com/deep-understanding-of-the-arima-model-d3f0751fc709

The most relevant part of the modelling set-up is the definition of the best parameters to be used, so several checks and analysis will be done:

- Dickey Fuller Test (from statsmodels package). This test is done to be done to identify if the series is non-stationary, if it is stationary, then the parameter d from ARIMA model should be "0".

  To identify the nature of data, it will be used the null hypothesis.

  H0: The null hypothesis: It is a statement about the population that either is believed to be true or is used to put forth an argument unless it can be shown to be incorrect beyond a reasonable doubt.

  H1: The alternative hypothesis: It is a claim about the population that is contradictory to H0 and what we conclude when we reject H0.

  H0: It is non-stationary

  H1: It is stationary

  We will be considering the null hypothesis that data is not stationary and the alternate hypothesis that data is stationary.

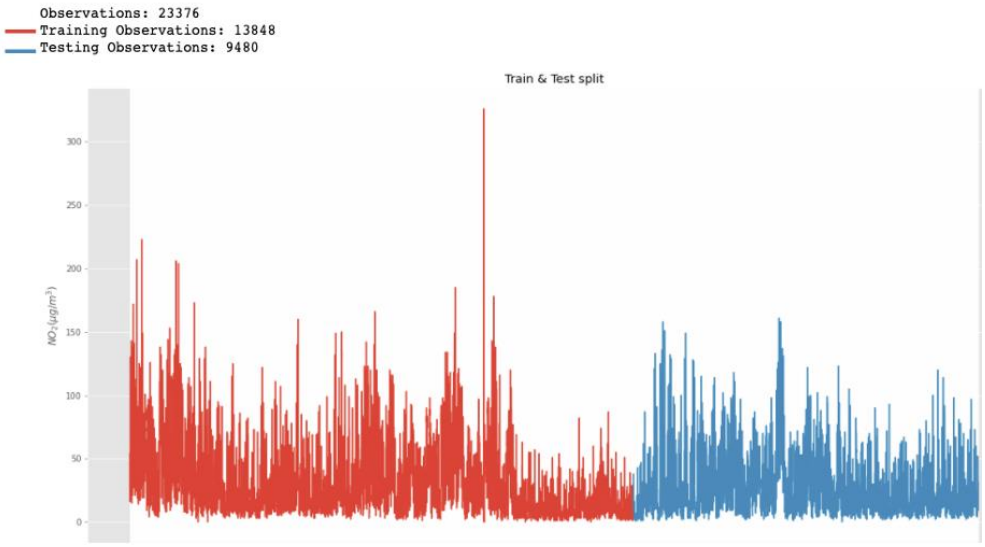
  This test is indicating that our data is stationary (a stationary process is a process whose mean, variance and autocorrelation structure don't change), so "d" parameter in ARIMA probably will be 0, :

```
Strong evidence against the null hypothesis(H0), reject the null hypothesis. Data is stationary
ADF Test Statistic : -9.656699472912415
p-value : 1.3879931518790027e-16
#Lags Used : 47
Number of Observations : 23328
```

  Test Dickey-Fuller result is validated, together the other parameters of a SARIMA model using an automatic function that validate each combination of parameters, using SARIMAX function from statsmodels, calculating the AIC (Akaike Information Criteria)[10] for each combination. The Akaike Information Criteria (AIC) is a widely used measure of a statistical model quantifying the goodness of fit and the simplicity of the model into a single figure. So, AIC is used to measure the relative quality of the model

[10] https://otexts.com/fpp2/arima-estimation.html

First, full data available needs to be split between train and test dataset:



As data is temporal series, this split needs to contain continuous data, not randomly split, the case of this study, it is 60%, from January 2019 to July 2020 (train dataset) and 40%, from August 2020 to August 2021 (test dataset).

Once, train-test data is split, an automatic function is run to identify the parameters that provide the lowest AIC value:

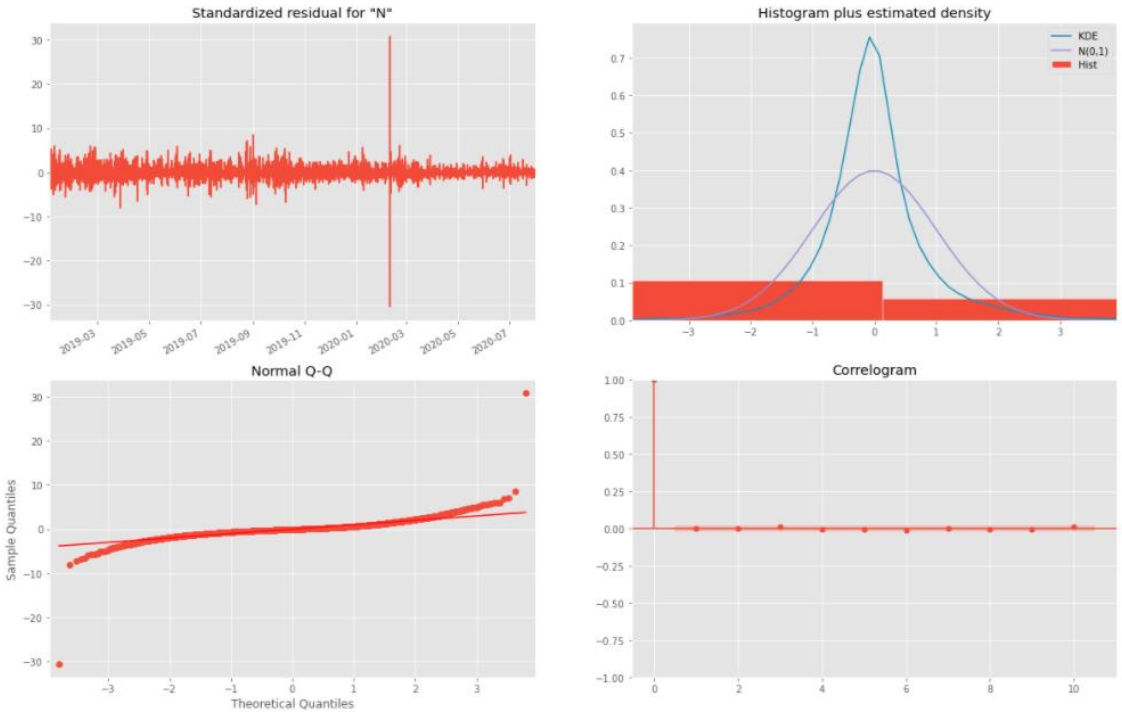
```
The smallest AIC is 103712.37525338921 for model SARIMAX(3, 0, 1)x(3, 1, 1, 12)
```

Based on above parameter the model is trained (using train data):

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                           NO2_index   No. Observations:                13848
Model:             SARIMAX(3, 0, 1)x(3, 1, 1, 12)   Log Likelihood              -51847.188
Date:                            Thu, 23 Dec 2021   AIC                         103712.375
Time:                                    16:03:14   BIC                         103780.165
Sample:                                01-01-2019   HQIC                        103734.959
                                     - 07-30-2020
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.9171      0.008    244.514      0.000       1.902       1.932
ar.L2         -1.0766      0.008   -143.275      0.000      -1.091      -1.062
ar.L3          0.1557      0.003     48.152      0.000       0.149       0.162
ma.L1         -0.9479      0.007   -126.997      0.000      -0.963      -0.933
ar.S.L12      -0.0690      0.008     -8.162      0.000      -0.086      -0.052
ar.S.L24       0.1103      0.004     24.972      0.000       0.102       0.119
ar.S.L36      -0.0711      0.007     -9.642      0.000      -0.086      -0.057
ma.S.L12      -0.9738      0.002   -468.017      0.000      -0.978      -0.970
sigma2       107.8414      0.236    457.600      0.000     107.380     108.303
===================================================================================
Ljung-Box (L1) (Q):                   0.10   Jarque-Bera (JB):          10244416.48
Prob(Q):                              0.75   Prob(JB):                         0.00
Heteroskedasticity (H):               0.69   Skew:                             0.40
Prob(H) (two-sided):                  0.00   Kurtosis:                       136.49
===================================================================================
```

The table in the middle is the coefficients table, where the values under “coefficients” are the weights of the respective terms.
Results are not bad, as p-value (P>[z] is less than 0.05.

In addition to this, to evaluate the model different diagnostics plots are used.



- Standardized residual. It has a uniform variance around mean of 0
- Histogram Estimated Density. Distribution looks like to be a normal distribution (with mean 0).
- Normal Q-Q. Most of the dots are falling with the red line, any significant deviations would imply the distribution is skewed.
- Correlogram. Plots show that residual errors are not autocorrelated, in case of autocorrelation would mean that there is some patter in the residual errors, which would not be explained by this model.

Model is not perfect, but it will be used in this study. After conclusions next steps will be analyzed trying to improve prediction.

Prediction in this study is based on 2 approaches:

- Static forecast: one step ahead forecast
- Dynamic forecasting for a prediction window of 2, 4, 6, 12 and 24 hours.

The firs method produce one-step ahead forecast, this will work much better than dynamic prediction into longer term, but also the added value of one-step ahead prediction is much less.

Different predictions have been evaluated using 3 metrics:

The **Mean absolute error** represents the average of the absolute difference between the actual and predicted values in the dataset. It measures the average of the residuals in the dataset.

The **Mean Squared Error** represents the average of the squared difference between the original and predicted values in the data set. It measures the variance of the residuals.

The **Root Mean Squared Error** is the square root of Mean Squared error. It measures the standard deviation of residuals.

Differences among these evaluation metrics

Mean Squared Error (MSE) and Root Mean Square Error penalizes the large prediction errors vi-a-vis Mean Absolute Error (MAE). However, RMSE is widely used than MSE to evaluate the performance of the regression model with other random models as it has the same units as the dependent variable (Y-axis).

MSE is a differentiable function that makes it easy to perform mathematical operations in comparison to a non-differentiable function like MAE. Therefore, in many models, RMSE is used as a default metric for calculating Loss Function despite being harder to interpret than MAE.

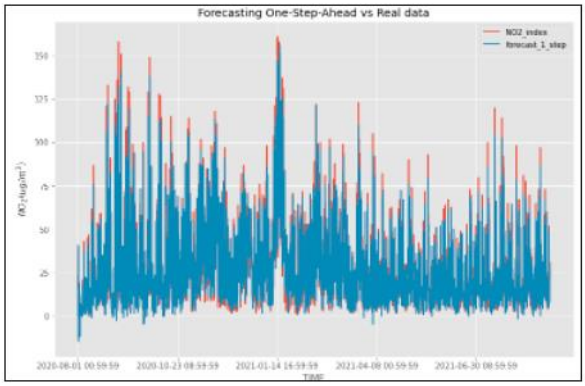
MAE is more robust to data with outliers.

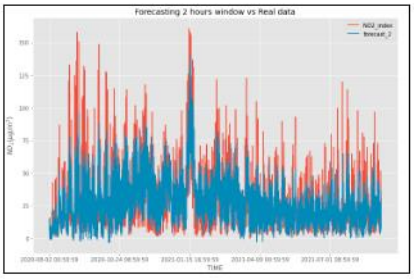
The lower value of MAE, MSE, and RMSE implies higher accuracy of a regression model. Although the parameters showed below are indicating that model is not so good. So in further phases of this study new models will need to be worked out.

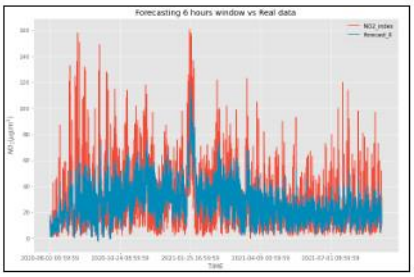
Prediction Models	MSE	MAE	RMSE
Static forecasting: one-step-ahead	5.8	78.6	8.8
Dynamic forecasting: 24 hour window	13.3	340.8	18.4
Dynamic forecasting: 12 hour window	12.7	316.1	17.7
Dynamic forecasting: 6 hour window	12.2	295.3	17.1
Dynamic forecasting: 2 hour window	10.2	219.6	14.8

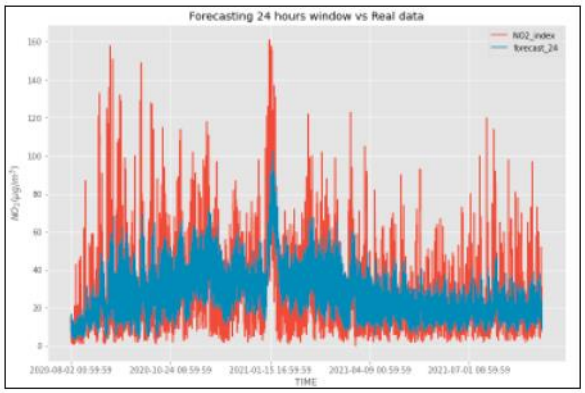
As expected, model works better with one-step ahead, although the value of this prediction is low, but prediction accuracy between 24h and 12 hours is very similar.

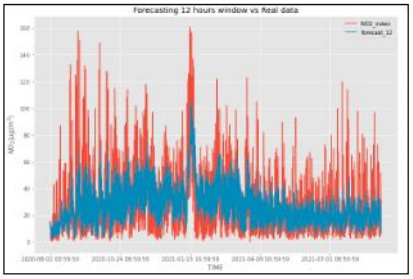
Comparation of prediction using the different methods (one-step-ahead, 2 hours, 4 hours, 6 hours, 12 hours, and 24 hours) vs the real data.











Although previous correlation matrix using the different weather variables available show a non-strong correlation, a test of the model has been done adding the two most correlated variables: Temperature & Wind Speed, although the model didn't improve:

MODEL EVALUATION with Exogeneus variables: WindSpeed & Temperature

Prediction Models	MSE	MAE	RMSE
Static forecasting: one-step-ahead	5.8	78.6	8.8
Static forecasting: one-step-ahead - ExogenousVariables (Temperature & WindSpeed)	5.8	78.6	8.8
Dynamic forecasting: 24 hour window	13.3	340.8	18.4
Dynamic forecasting: 24 hour window - ExogenousVariables (Temperature & WindSpeed)	13.2	340.1	18.4
Dynamic forecasting: 12 hour window	12.7	316.1	17.7
Dynamic forecasting: 12 hour window - ExogenousVariables (Temperature & WindSpeed)	12.7	315.4	17.7
Dynamic forecasting: 6 hour window	12.2	295.3	17.1
Dynamic forecasting: 6 hour window - ExogenousVariables (Temperature & WindSpeed)	10.2	295.3	17.1
Dynamic forecasting: 2 hour window	10.2	219.6	14.8
Dynamic forecasting: 6 hour window - ExogenousVariables (Temperature & WindSpeed)	10.2	219.6	14.8

LIVE DATA PREDICTION & VISUALIZATION

Once a good enough model is ready, now a prediction can be worked out.

For a useful use of this study, the prediction will be done based on real time data provided by "Ayuntamiento de Madrid" for every Air Station data and provided in a web interface for a friendly visualization and use.

Ayuntamiento de Madrid in its web [11] about Air Quality, provide measurements for all Air stations updated in real time, being the data updated every hour. This data update will happen between the minutes 20 and 30 of every hour.

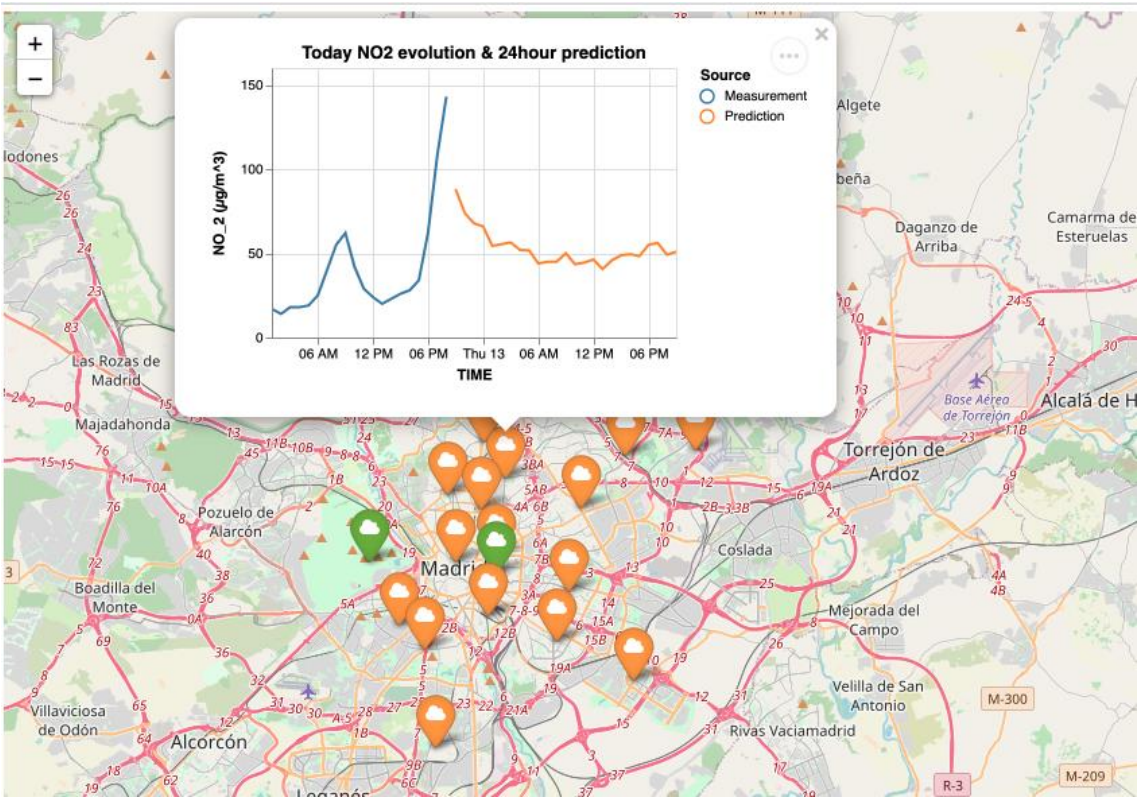
(Disclaimer of live data: This data is live data automatically measured by the Air Stations without any quality check, so data is pending to be reviewed and validated)

There are different considerations taken during the realization of this part:

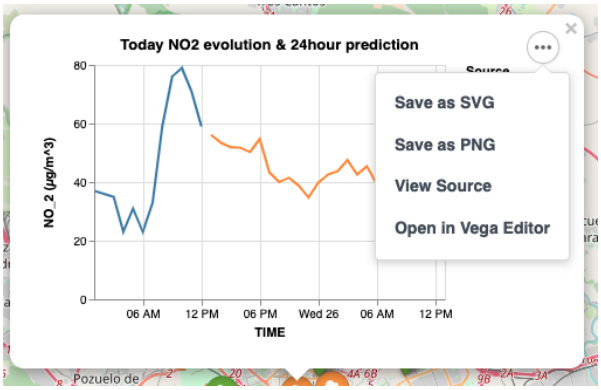
- Real time data reading, prediction and visualization is done for every Air Station.
- The real time component adds some complexity that need to be considered to guarantee that complete process works:
    - Create a lag time between data available and prediction due to publication time.
    - 24 new ROWs need to be added in the data frame to allocate prediction based on current time of the prediction.
- Live data is in the same format than historical data, so the same transformation than in data cleaning and transformation need to be done (same code)
- Live data provided contains only the measurements related to the current date.
- Prediction with so few records (only the ones provided with the real time data) is not enough to work a good prediction, so all historical data used during model creation is added to the data of every station, this makes that model works better.

11 https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=41e01e007c9db410VgnVCM2000000c205a0aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextfmt=default

For the data visualization folium[12] library has been used, this library allows us a geographical representation of all Air Station, together with a chart adding measurements and predictions for the next 24 hours for the related station. This map is saved in html format to be used in the web application

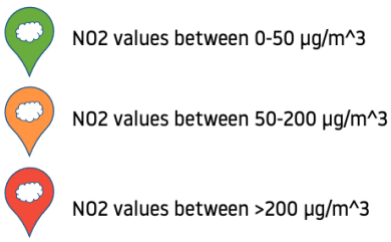


Measurements/Predictions charts can be exported clicking in the right upper corner



[12] https://python-visualization.github.io/folium/

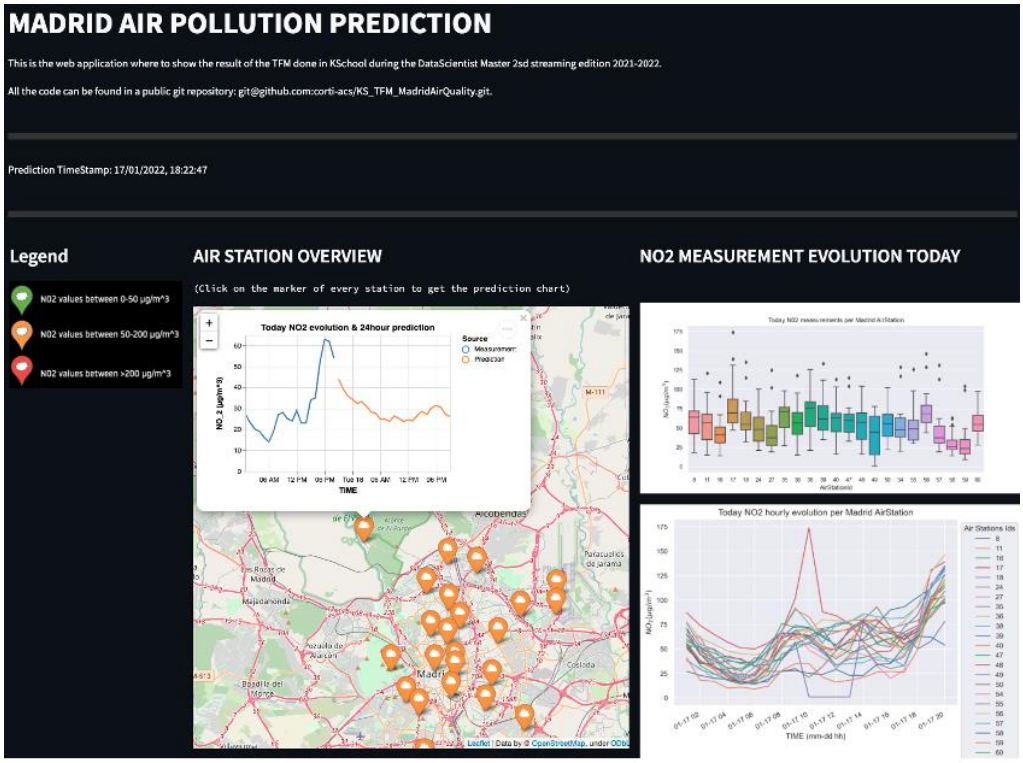
The markers have different color to just have a quick view of NO2 values in every station:



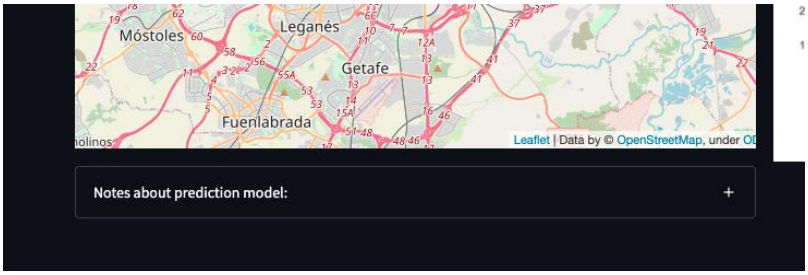
Web application also show some charts to help to the dashboard user to have more information about measurement evolution in the day for every Air Station:

NO2 values during the prediction date for every Air Station
NO2 distribution in a box plot to easily compare the values in every Air Station

Web App view:



Below the map, there is an expanded section (“Notes about prediction model”) where the model details are explained for a better understanding of the model used and prediction quality.



5. CONCLUSIONS & NEXT STEPS

From this study, there are different outcomes can be taken away.

Data series

- **Special data events.** WHO (World Health Organization, aka OMS) define as unhealthy when NO2 values are higher than 200 µg/m^3. These event in Madrid are special events in the data series, having short peaks. (141 measurements registered from all stations with higher values than 200 µg/m^3 from half million measurements), What is reflecting an unbalanced data series.

- **Some geographical data observation:** Although the average values in Madrid remain below 50 µg/m^3 the most of the time, the unhealthy limit is overpass in the most of Air Stations several times in a year, being the Air Station located in "Plaza Elíptica" the station where more often this limit is overpassed and areas like " La Casa Campo" and "El Pardo" the areas with less pollution (as it could be expected).

- **Patterns**: In the data series different patterns can be observed
    - Hourly: Two pollution peaks around 8:00 am and 21:00 pm
    - Weekly day: Lowers values on the weekends, keeping stable values, being Tuesdays the days with higher pollution values
    - Monthly: Higher pollution values in winter and autumn months.

- **Lag data correlation** with 1 step (1 hour in this study) is visible plotting lag plot, but this correlation is quickly lost when we add more than 1 step…

- **Correlation of NO2 vs weather parameter**. N02 values show some correlation with windspeed, although calculating Pearson correlation coefficient there is not so high correlation how it could be expected.

  Other data observation:
    - Seasonality. Clear seasonality component
    - Trend. Slight negative trend, probably due to COVID restriction impact in 2020/2021 vs 2019
    - Noise. Relevant amount of noise, what can indicate that pollution increases are happening randomly when different conditions push the NO2 levels up.
    - Stationary

Modeling and Prediction

Current model is very conservative, having difficulties to predict special events, like "unexpected" peaks, probably due to these events are special events, happening occasionally and model is not capable to identify.

The help of Exogenous variables like Wind Speed or Temperature should have helped to the model prediction, but the model evaluation metrics indicated that it is not the case.

Next steps to be done to improve the study:

- Use another model different than SARIMA. This model has limitations, being hard to model the nonlinear relationship between variables. There are other regression models like XGBoost that could work better or maybe go for something more complex like Convolutional Neural Network (CNN)[13], where air quality and weather can be interpolated giving a more robust model.
- Create a different model for every station. This could improve the details of the prediction.
- Add other component to the model like Traffic index, this could be used as a new endogen variable in the model. This data is available hourly, so it should be possible to link with the current data used in this model.

6. ADDENDUM

WEB APPLICATION INSTRUCTIONS

In this section it is explained the steps to follow to run web interface.

In this case, Streamlit[14] library has been used; for a proper use it is recommended to follow the next steps:

a. From "webapp" folder in the repository execute in the console "LiveDataPrediction.py". This will read live data and will do the prediction. This will take around 5 minutes.
b. In the same folder, execute in the console the WebApp.py using this command: "streamlit *run WebApp.py*", this will launch the web application.

Note: Streamlit run WebApp.py can be executed without the need to execute the prediction, web app will use the last prediction done in the past and saved in html file, so if the first step is not executed is important to verify the "Prediction Time Stamp" to know the hour and date when the prediction was calculated. This can be verified by a Time Stamp in the Dashboard.

[13] https://www.nature.com/articles/s41598-021-91253-9

[14] https://streamlit.io/

AIR STATION NETWORK.

STATIONS CODE

Stations with asterisk (*) changed the id from the date indicated, just for the adaptation to the national identification system related to AirQuality data interchange.

28079001	Pº. Recoletos	Baja.- 04/05/2009 (14:00 h.)
28079002	Glta. de Carlos V	Baja.- 04/12/2006 (11:00 h.)
28079003 28079035(*)	Pza. del Carmen	* Código desde enero 2011
28079004	Pza. de España	
28079005 28079039(*)	Barrio del Pilar	* Código desde enero 2011
28079006	Pza. Dr. Marañón	Baja.- 27/11/2009 (08:00 h.)
28079007	Pza. M. de Salamanca	Baja.- 30/12/2009 (14:00 h.)
28079008	Escuelas Aguirre	
28079009	Pza. Luca de Tena	Baja.- 07/12/2009 (08:00 h.)
28079010 28079038(*)	Cuatro Caminos	* Código desde enero 2011
28079011	Av. Ramón y Cajal	
28079012	Pza. Manuel Becerra	Baja.- 30/12/2009 (14:00 h.)
28079013 28079040(*)	Vallecas	* Código desde enero 2011
28079014	Pza. Fdez. Ladreda	Baja.- 02/12/2009 (09:00 h.)
28079015	Pza. Castilla	Baja.- 17/10/2008 (11:00 h.)
28079016	Arturo Soria	
28079017	Villaverde Alto	
28079018	C/ Farolillo	
28079019	Huerta Castañeda	Baja.- 30/12/2009 (13:00 h.)
28079020 28079036(*)	Moratalaz	* Código desde enero 2011
28079021	Pza. Cristo Rey	Baja.- 04/12/2009 (14:00 h.)
28079022	Pº. Pontones	Baja.- 20/11/2009 (10:00 h.)
28079023	Final C/ Alcalá	Baja.- 30/12/2009 (14:00 h.)
28079024	Casa de Campo	
28079025	Santa Eugenia	Baja.- 16/11/2009 (10:00 h.)
28079026	Urb. Embajada (Barajas)	Baja.- 11/01/2010 (09:00 h.)
28079027	Barajas	
28079047	Méndez Álvaro	Alta.- 21/12/2009 (00:00 h.)
28079048	Pº. Castellana	Alta.- 01/06/2010 (00:00 h.)
28079049	Retiro	Alta.- 01/01/2010 (00:00 h.)
28079050	Pza. Castilla	Alta.- 08/02/2010 (00:00 h.)
28079054	Ensanche Vallecas	Alta.- 11/12/2009 (00:00 h.)
28079055	Urb. Embajada (Barajas)	Alta.- 20/01/2010 (15:00 h.)
28079056	Plaza Elíptica	Alta.- 18/01/2010 (12:00 h.)
28079057	Sanchinarro	Alta.- 24/11/2009 (00:00 h.)
28079058	El Pardo	Alta.- 30/11/2009 (13:00 h.)
28079059	Parque Juan Carlos I	Alta.- 14/12/2009 (00:00 h.)
28079086 28079060(*)	Tres Olivos	Alta.- 14/01/2010 (13:00 h.) * Código desde enero 2011

PARAMETERS, UNITS & MEASUREMENT TECHNICS

Magnitud		Abreviatura o fórmula	Unidad medida	Técnica de medida	
01	Dióxido de Azufre	SO_2	µg/m ^{3}	38	Fluorescencia ultravioleta
06	Monóxido de Carbono	CO	mg/m^3	48	Absorción infrarroja
07	Monóxido de Nitrógeno	NO	µg/m^3	08	Quimioluminiscencia
08	Dióxido de Nitrógeno	NO_2	µg/m ^{3}	08	Id.
09	Partículas < 2.5 µm	PM2.5	µg/m^3	47	Microbalanza
10	Partículas < 10 µm	PM10	µg/m^3	47	Id.
12	Óxidos de Nitrógeno	NOx	µg/m^3	08	Quimioluminiscencia
14	Ozono	O_3	µg/m ^{3}	06	Absorción ultravioleta
20	Tolueno	TOL	µg/m^3	59	Cromatografía de gases
30	Benceno	BEN	µg/m^3	59	Id.
35	Etilbenceno	EBE	µg/m^3	59	Id.
37	Metaxileno	MXY	µg/m^3	59	Id.
38	Paraxileno	PXY	µg/m^3	59	Id.
39	Ortoxileno	OXY	µg/m^3	59	Id.
42	Hidrocarburos totales (hexano)	TCH	mg/m^3	02	Ionización de llama
43	Metano	CH4	mg/m^3	02	Id.

Important Note: Not all stations are gathering new measurements for all parameters, due to technical reasons any station can stop the measurements.

WEATHER STATION NETWORK.

STATIONS CODE LIST

CÓDIGO	ESTACIÓN
28079102	J.M.D. Moratalaz
28079103	J.M.D. Villaverde
28079104	E.D.A.R. La China
28079106	Centro Mpal. De Acústica
28079107	J.M.D. Hortaleza
28079108	Peñagrande
28079109	J.M.D.Chamberí
28079110	J.M.D.Centro
28079111	J.M.D.Chamartin
28079112	J.M.D.Vallecas 1
28079113	J.M.D.Vallecas 2
28079114	Matadero 01
28079115	Matadero 02
28079004	Plaza España
28079008	Escuelas Aguirre
28079016	Arturo Soria
28079018	Farolillo
28079024	Casa de Campo
28079035	Plaza del Carmen
28079036	Moratalaz
28079038	Cuatro Caminos
28079039	Barrio del Pilar
28079054	Ensanche de Vallecas
28079056	Plaza Elíptica
28079058	El Pardo
28079059	Juan Carlos I

PARAMETERS, UNITS & MEASUREMENT TECHNICS

CÓDIGO	PARÁMETRO	UNIDAD DE MEDIDA	TÉCNICA DE MEDIDA
80	RADIACIÓN ULTRAVIOLETA	Mw/m2	98
81	VELOCIDAD VIENTO	m/s	98
82	DIR. DE VIENTO	-	98
83	TEMPERATURA	ºC	98
86	HUMEDAD RELATIVA	%	98
87	PRESION BARIOMETRICA	mb	98
88	RADIACION SOLAR	W/m2	98
89	PRECIPITACIÓN	l/m2	98