Exploring the relationship between pollution and weather conditions: an European analysis

Andrea D'Amicis, Gaia Gubelli, and Emmanuele Lotano
Data Management project
(Dated: Academic year 2023/2024)

INDEX

INTRODUCTION

In recent years, the issue of environmental pollution has become one of the most pressing problems that global society faces. Mainly air pollution has significant implications on human health and climate change.

This study aims to provide a descriptive overview of the dynamics governing air pollution in relation to weather conditions through a data collection analysis. The relationship between pollution and weather conditions will be explored and compared, with a particular focus on their impact on human health.

Research questions

The retrieval of the data was guided by some research questions, formulated based on some theoretical aspects[1][2][3].

Air pollutants typically occur in two physical states: solid particulate matter (such as PM10 or PM2.5) and pollutants in the gas phase (such as O3, CO, NO2, SO2). With this initial consideration, we came up with the following research questions.

- **European descriptive analysis**:

  1. *What are the European average O3 levels across maximum temperature ranges?*

     Sunny days and high temperatures can have a negative impact on air quality. When temperatures are above average, the resulting heat waves can make the air we breathe dangerous. Nitrogen oxide (denoted as NOx) emissions from various sources such as industry, vehicle traffic and petrol vapours reacts with sunlight to form ozone (denoted as O3). Warm weather is a significant factor in ozone production, therefore countries with a warm and sunny climate may face days with intense smog episodes.

  2. *What are the European average NO2 levels across precipitation ranges?*

     Rain plays an important role in lowering concentrations of pollutants in the air. When raindrops fall, they can agglomerate gaseous pollutant particles (such as NO2 and SO2) in a process known as "coagulation". These pollutants are generated through the combustion of fossil fuels, industrial processes and vehicles emissions. The phenomenon of the "coagulation" allows pollutants spread in the atmosphere to be captured and precipitated to the ground, thus contributing to the cleanliness of the air. It is interesting to note that rain (in particular during moderate rainfall) is more effective than snow in this process of atmospheric purification.

3. *What are the European average PM10 levels across different wind speed ranges?*
4. *What are the European average values of PM10 by season?*

The acronym "PM" stands for "Particulate Matter", followed by the maximum particle diameter in micrometres. These particulate matter are mainly divided into two categories: PM10, which includes coarser particles such as dust, pollen and spores, and PM2.5, consisting of ultrafine particles mainly originated from industrial and vehicular emissions. Larger particles are retained by the respiratory system before reaching the lungs, while ultrafine particles can enter the alveoli and even the bloodstream, causing inflammation and respiratory problems.
Wind can play a beneficial role in dispersing pollutant particles in the air. When pollution accumulates in a region, wind can help disperse the pollutants, thus reducing pollutant concentrations in the affected area. However, the geography and morphology of the terrain can occasionally pose a challenge to the wind in dispersing pollutants. Furthermore, in case of forest fires, large quantities of pollutants can be transported far from their place of origin, polluting areas far and near.
Moreover, during periods of low temperatures (especially in winter and spring), there is an increase in pollution levels compared to other seasons. This is because the colder air tends to form a warmer upper layer, acting as a kind of lid that prevents the troposphere from disposing of pollutants in the higher zones. Moreover, during these seasons, the use of fossil fuels for domestic heating increases, further contributing to higher pollution levels.

- **Italian descriptive analysis**:

1. *How do pollution measures vary over time in Italy?*

Having made the previous considerations for European countries, we focused on our home country. We decided to analyse the monthly average levels of NO2, O3 and PM2.5 over the year 2022.

2. *What are the 5 records with the highest PM2.5 values observed in Italy? Where were these values observed? What were the meteorological conditions on those days?*

DATA ACQUISITION

For the aims of our research, we obtained the necessary data by querying two APIs.

Pollution

The first API, concerning air pollution parameters, is the following: `https://openaq.org`. To use this API, it was necessary to obtain the API key. In order to retrieve a dataset containing daily measurements for the entire year 2022, the API was queried to obtain the measurements of each city in the main European states, depending on the availability of the API.
The parameters requested for each call are as follows:

- PM10: Particulate Matter with a diameter less than or equal to 10 μm;
- PM2.5: Particulate Matter with a diameter less than or equal to 2.5 μm;
- O3: Ozone;
- NO2: Nitrogen dioxide;
- NO: Nitric oxide;
- CO: Carbon monoxide;
- NOx: Nitrogen oxide;
- SO2: Sulfur dioxide.

All parameters are expressed according to the unit of measurement $\mu g/m^3$. The data acquisition algorithm was designed and structured to download as much data as possible by exploiting the structure of the API itself and the endpoints it offers. As a first step, the list of cities provided by the API was obtained for each European state (uniquely identified by an ISO code).
The states used in the analysis are:

- Italy (IT)
- Switzerland (CH)
- Spain (ES)
- France (FR)
- Belgium (BE)
- Netherlands (NL)
- Germany (DE)
- Portugal (PT)
- Great Britain (GB)
- Ireland (IE)

- Austria (AT)
- Norway (NO)
- Finland (FI)
- Sweden (SE)

The response of each call contains the daily measurements with an hourly time granularity. However, in order to have a daily granularity, the hourly detections were averaged, thus aggregating them into a single daily value. In addition, the geographical coordinates of the detection are associated with each city.

Weather

In order to integrate the first dataset with daily meteorological conditions, the open source API `https://open-meteo.com` was used for each previously obtained record. The implemented algorithm retrieves the data from the geographical coordinates of previous pollution records. Each request to the API provides detections for the following parameters:

- WMO code
- Max. temperature ($^{\circ}C$)
- Min. temperature ($^{\circ}C$)
- Wind Speed (km/h)
- Apparent max. temperature ($^{\circ}C$)
- Apparent min. temperature ($^{\circ}C$)
- Precipitation (sum in mm)
- Rain (sum in mm)
- Snow (sum in mm)

Storage

We decided to store the two raw datasets in a document-oriented database and in two separate collections, in particular using the database management system MongoDB. Our choice was guided by the various advantages that such model offers: the documents are structured in a BSON format, a schema that facilitates the integration of the data at a later stage as the fields can themselves have sub-documents; furthermore, this model does not impose a schema on the data, allowing for greater flexibility and efficiency when dealing with large data sets.


```
_id: ObjectId('6585af1b96e7d3da98ec32f2')
State : "CH"
City : "Basel-Landschaft"
Date : "2022-01-11"
Latitude : 47.5410842894654
Longitude : 7.583269599999999
Pm10 : 2.804947826086957
Pm25 : 2.707542028985508
O3 : 2.609584057971014
No2 : 6.032176086956522
No : 0
Co : 0
Nox : 0
So2 : 0.2206195652173913
```


FIG. 1: Example of a document of the Pollution collection (raw).


```
_id: ObjectId('6585b6a7532eb78cb75582d9')
State : "CH"
City : "Basel-Landschaft"
Date : "2022-01-11"
Latitude : 47.5410842894654
Longitude : 7.5832695999999995
WMO_code : 3
TemperatureMin : -2.5
TemperatureMax : 3.4
WindSpeed : 10.8
ApparentTMAX : 0.4
ApparentTMIN : -5.7
PrecipitationSum : 0
RainSum : 0
SnowfallSum : 0
```


FIG. 2: Example of a document of the Weather collection (raw).

At the end of this phase we obtained 180'768 documents for each collection.

DATA PROFILING

In order to assess the quality of the data obtained, the two dimensions that seemed most appropriate were analysed, separately for the two collections. The first dimension tested was completeness, assessed both at the attribute level (by evaluating the percentage of missing values for each attribute) and at the collection level (by evaluating the presence of missing values in the two collections overall).
For the second dimension it was decided to analyse the consistency of the data.

Completeness

Pollution
Analysing the pollution dataset, we noticed that the

values represented by zeros were actually missing values where the API did not give an answer. This is due to the fact that by averaging to get a single daily value, the algorithm gave zero average when it was actually not possible to calculate it precisely. This happened when, for that particular day and place, the API did not provide any data in response. Consequently, to assess the completeness of the pollution dataset, we considered and evaluated both NaNs and zeros as they conceptually indicate the same thing.

	TOT	PM10	PM2.5	O3	NO2	NO	CO	NOx	SO2
NaN	224188	22342	22342	22342	22342	22342	22342	22342	22342
% NAN	8	12	12	12	12	12	12	12	12
0	705612	45125	71175	42973	15750	156836	119825	158426	95502
% 0	27	24	39	23	8	86	66	87	52

TAB. I: Completeness measures for Pollution.

As far as table completeness is concerned, the dataset presents 27% of 0 and 8% of NaN (as we can see in the first column of TAB. I), thus having 35% missing values. Focusing instead on attribute completeness, it emerges that the attributes “NOx”, “CO” and “NO” present almost all missing values in the observations. For the other parameters, however, the percentage of missing values is more moderate. As mentioned above, there are 22’342 occurrences in the dataset whose parameter measurements are completely null.

Weather

As opposite to the pollution dataset, here the completeness was evaluated with a different semantics. Here the zeros represent correct measurements, not resulting from a calculation but from the simple GET request of each call, indicating the absence of a certain weather phenomenon. For example, for attributes such as rainfall measurement or temperature, it is plausible to observe a value equal to zero. Therefore, in order to correctly assess completeness for this dataset, we only considered NaNs and obtained the following results.

	TOT	TempMin	TempMax	AppTempMin	AppTempMax
NaN	45183	7	7	7	7
% NaN	1	0	0	0	0

	WMOCode	WindSpeed	PrecSum	RainSum	SnowSum
NaN	7	7	7	7	7
% NaN	0	0	0	0	0

TAB. II: Completeness measures for Weather.

The dataset contains 45’183 missing values, which is less than 1% of the entire collection. If we look at the missing values of the meteorological parameters, however, the number of missing values is minimal. This stems from the fact that during the acquisition phase, if the longitude and latitude had zero values in the pollution dataset, then the call returned the atmospheric values from the previous call. This does not, however, cause a problem since during the cleaning and integration phase, these documents will be removed.

Consistency

Pollution

Looking at the ranges of values assumed by the pollution attributes, we wanted to assess their consistency with the values observed in the reality and the presence of any outliers. For example, we note that in almost all variables, the minimum value assumed is negative. In addition, for some variables the maximum value assumed is improbable. This is not likely and it is probably due to measurement errors, or non-natural phenomena such as a fire.

	PM10	PM2.5	O3	NO2	NO	CO	NOx	SO2
Min	-333	-499	-249	-3333	-249	-200	0	-206
Max	204	85	70	60	30	73131	0	37

TAB. III: Consistency measures for Pollution.

Weather

As done for the first dataset, looking at the ranges we noticed inconsistencies. As can be seen in the TAB. IV, the variables *WindSpeed* and *PrecipitationSum* show negative minimum values of -3 and -11 respectively. Furthermore, the maximum value assumed by the variable *TemperatureMin* is 73°C, which is significantly higher than the maximum value assumed by the variable *TemperatureMax*, which is 48°C. This is inconsistent with the nature of the variables themselves.

To investigate this problem further, we performed queries on the database. It turned out that there are some documents whose measurements are inconsistent. For example, the document for the city of Reggio Emilia on 13-12-2022 is shown in the FIG. 3. It assumes improbable values in the parameters concerning temperature, wind speed and precipitation.

	TempMin	TempMax	AppTempMin	AppTempMax
Min	-38	-29	-42	-34
Max	73	48	44	48

	WMOCode	WindSpeed	PrecSum	RainSum	SnowSum
Min	0	-3	-11	0	0
Max	75	76	106	106	59

TAB. IV: Consistency measures for Weather.

```
_id: ObjectId('6585ad8b96e7d3da98ec1898')
State : "IT"
City : "Reggio Nell'Emilia"
Date : "2022-12-13"
Latitude : 44.299722
Longitude : 10.430278
WMO_code : 53
TemperatureMin : 73
TemperatureMax : -4.4
WindSpeed : -2.9
ApparentTMAX : 6.6
ApparentTMIN : -6.1
PrecipitationSum : -7.8
RainSum : 3.6
SnowfallSum : 0.1
```

FIG. 3: Example of inconsistent values in a document (Weather).

DATA INTEGRATION AND DATA CLEANING

Firstly, we combined the two datasets containing the results of pollution and meteorological conditions in such a way as to have, for each day and for each city, a single, coherent view of the observations. Once this was done, we cleaned up the dataset by handling the issues that arose during the data quality assessment.

Pollution
As a first step, we removed all those observations in which the parameters had negative values or excessively high values. Furthermore, as already noted, the variables CO, NOx and NO contain almost all missing values and we therefore decided to remove them completely from the analysis. For those documents that still have missing values after the previous steps (representing a small part of the data), we decided to carry out the mean replacement, for each European state taken individually, in order to ensure the highest possible consistency and reliability.

Weather
Once the pollution issues were resolved, the only issue to be fixed for the meteorological attributes concerned inconsistent and improbable measurements. Based on what emerged from the data quality assessment, we decided to remove those documents that had one or more of the following conditions:

1. Min. temperature > Max. temperature

2. Apparent min. temperature > Apparent max. temperature

3. Min. temperature > 50 OR Apparent min. temperature > 50

There were 617 documents fulfilling the first or second condition, and 285 documents fulfilling the third condition.

Based on the verifications carried out by querying the database, we noticed that all those documents with negative values in the attributes *WindSpeed* and *Precipitation Sum* fell into this category.

DATA ENRICHMENT

At this stage, we wanted to enrich the final dataset by adding additional information from external sources. To provide greater interpretability to the WMO code variable, we created an additional field containing the code description, obtained from the following website: https://www.nodc.noaa.gov/archive/arc0021/0002199/1.1/data/0-data/HTML/WMO-CODE/WMO4677.HTM[4].

In order to facilitate an easier handling of the search questions, we completed the data by adding two attributes:

- *Season*, which are Winter, Summer, Spring, Autumn;

- *Region*, which divides the European states into "Northern Europe", "Western Europe" and "Southern Europe" according to United Nations Geoscheme for Europe[5].

At the end of this phase, we obtained a collection containing 99'048 documents.

FINAL STORAGE

Once the data enrichment phase was performed, we created three new collections with the cleaned data:

- *collectionMerged*, containing the aggregated data;

- *collectionPollutionCleaned*, containing only the final pollution data;

- *collectionWeatherCleaned*, containing only the final weather data.

In order to facilitate our queries, we decided to create objects in each document as shown in FIG. 4.

The final structure contains all the necessary information, such as state, city, date and geographical coordinates. Pollution and weather measurements are clearly visible, accompanying each value with its unit of measurement.

```
_id: ObjectId('65ce383f5b493dc1b735ba1d')
State : "CH"
City : "Basel-Landschaft"
Date : 2022-01-11T00:00:00.000+00:00
Latitude : 47.5410842894654
Longitude : 7.583269599999999
▾ Measurements : Object
    Pm10 : 2.804947826086957
    Pm25 : 2.707542028985508
    O3 : 2.609584057971014
    No2 : 6.032176086956522
    So2 : 0.2206195652173913
    Unit : "µg/m³"
▾ WMO : Object
    Code : 3
    Description : "Clouds generally forming or developing"
▾ Temperature : Object
    Min : -2.5
    Max : 3.4
    Unit : "°C"
▾ ApparentTemperature : Object
    Min : -5.7
    Max : 0.4
    Unit : "°C"
▾ WindSpeed : Object
    Value : 10.8
    Unit : "km/h"
▾ Precipitation : Object
    Sum : 0
    Rain : 0
    Snowfall : 0
    Unit : "mm"
Region : "Western Europe"
Season : "Winter"
```

FIG. 4: Final structure of the documents in the integrated collection.

QUERIES

Once we had obtained a database stored in MongoDB that was complete and cleaned, we focused on answering our research questions, formulated in the first phase of the analysis.

The research questions that follow are intended to be a simple exploratory analysis of the dataset obtained. It is important to remember the complexity of the interaction between the different pollution and meteorological parameters, which does not allow with simple plots to represent the real relationship between them without an in-depth study.

1. European average O3 levels over temperature

To answer our first research question, we ran a query which, after discretizing the Max. temperature variable into classes, returns as output a DataFrame that for each class and for each geographical region calculates the average O3 level.

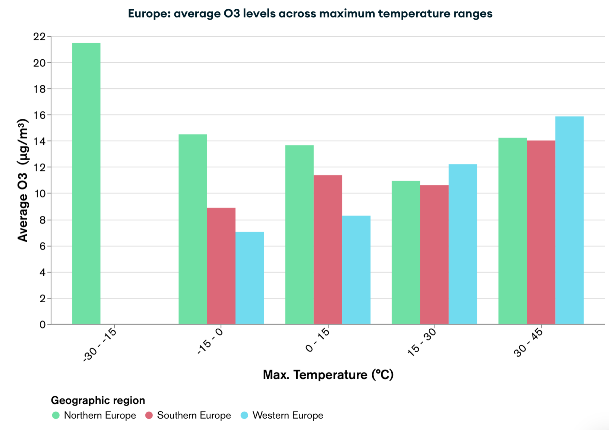

FIG. 5: Europe: average O3 levels across maximum temperature ranges.

Plotting the results in a bar graph, as we can see in FIG. 5, it shows that the trend in O3 levels reflect the theoretical considerations made initially, with the exception of the Northern European region. In fact, an increasing trend in O3 levels can be seen as temperatures rise. One data point that stands out is the value for the temperature range [-30, -15] for Northern Europe. By interrogating the database, we were able to see that there are 18 documents relating only to the Scandinavian countries where extreme temperatures are observed. A possible reason for high O3 levels can be found in the phenomenon of "temperature inversion", as mentioned in the theoretical introduction.

2. European average NO2 levels over daily rain

As for the second research question, we ran a query that, after discretizing the variable related to the daily rain measurement into classes, returns as output a DataFrame that for each class and each geographical region calculates the average NO2 level.

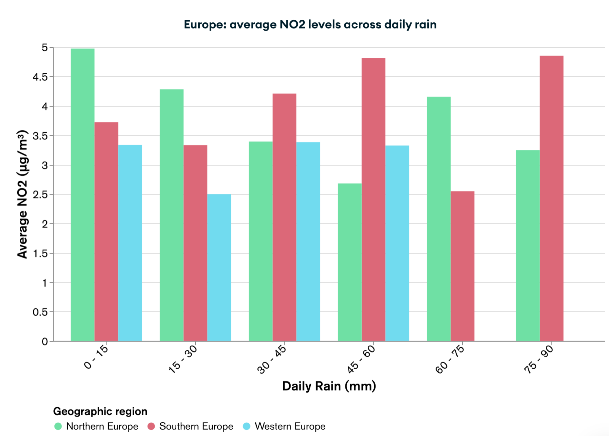

FIG. 6: Europe: average NO2 levels across daily rain.

From the FIG. 6 we can see that there is no clearly defined trend. In fact, as the sum of daily rainfall millimetres increases, there is no average decrease in the NO2 level. This is a clear example of how complex the phenomenon in issue is: as mentioned earlier, only weak and continuous rainfall can lead to an improvement in air quality. Moreover, the days with the best air cleanliness levels are probably observed on the days following the rainy days and not on the days themselves.

During the construction of the graph, class [90,105] and class [105+] were represented by only 2 documents each and we therefore preferred to exclude them from the analysis.

3. European average PM10 levels over wind speed

Likewise for the first two research questions, after discretizing the wind speed variable, the average PM10 was calculated for each class and geographical region.

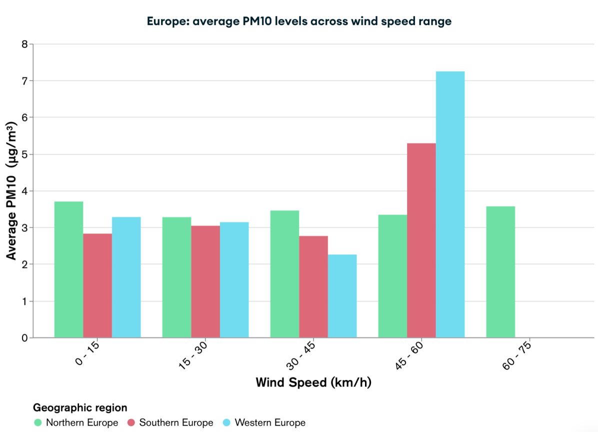

FIG. 7: Europe: average PM10 levels across wind speed range.

As can be seen from FIG. 7, as far as the first three classes are concerned, they are almost homogeneous and we cannot assume that a trend is present. On the other hand, for class [45,60], there is a clear difference in the increase of the average PM10 value in the presence of stronger winds, particularly for Western Europe. This increase could be traced back to the fact that in the presence of strong winds, turbulent motions are able to transport large quantities of pollutants to the surroundings, preventing dispersion in the atmosphere and therefore not leading to an effective improvement in air quality.

As we have done before, during the construction of the graph, class [75,90] was represented by only 2 documents and we therefore preferred to exclude it from the analysis.

4. European average PM10 levels across seasons

It was also our intention to assess the average PM10 levels for each season of the year. In this way, the considerations made at the beginning can be observed in FIG. 8.

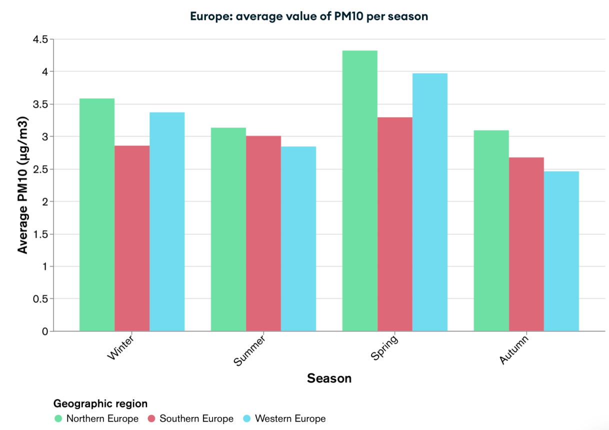

FIG. 8: Europe: average value of PM10 per season.

The highest average levels are found in the spring, probably due to the fact that PM10 is mainly composed of dust, pollen and spores. As one might expect, following the spring, the winter also shows slightly higher values, with the exception of the Southern Europe regions.

5. Italian pollution measures over time

One of our questions of interest was to observe the monthly trend of Italian pollution values taken individually.

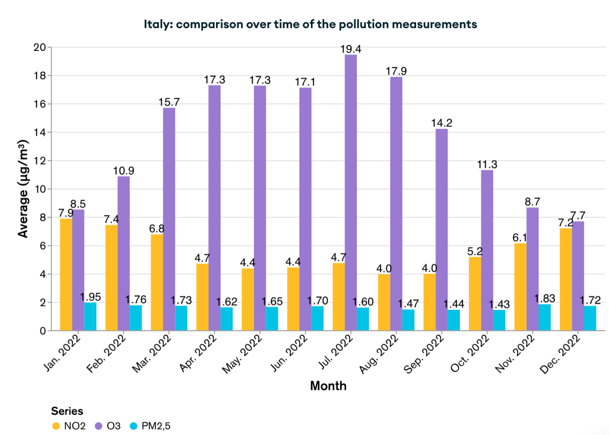

FIG. 9: Italy: comparison over time of the pollution measurements.

As can be seen from FIG. 9, the considerations made are as follows:

- NO2: as this is a compound of gaseous polluting particles produced by industrial processes, vehicles and domestic heating, it has higher values during the colder months (e.g. January and February) and lower values during the warmer months (e.g. August and September);
- O3: in this case, an opposite trend to the previous case can be observed where maximum ozone levels are reached in the summer months. This is probably due to the fact that ozone is formed on the sunniest days with high temperatures;
- PM2.5: the average level of this pollutant remains more or less constant throughout the year. However, slightly higher values occur in the colder months due to the probable "temperature inversion".

6. Highest PM2.5 values observed in Italy

City	Date	PM2.5	WindSpeed	TempMin	TempMax	PrecSum
Brindisi	22-11	84.8	43.4	10.2	19.5	6.7
Brindisi	21-01	73.7	22.7	7.5	12.6	3.3
Brindisi	09-02	66.6	22.1	6.3	15.3	0
Brindisi	19-01	66.5	13.9	2.2	11.3	0
Brindisi	01-11	66.0	14.2	16.2	25.4	0

TAB. V: Highest PM2.5 values observed in Italy.

With an interest in finding out which observations were the most polluted in Italy, as far as PM2.5 is concerned, it emerged that all five were in the city of Brindisi. We can see that PM2.5 levels are higher than the averages represented in FIG. 9 and these appear to be a problem for people's health, with reference to the Air Quality Index whose ranges can be seen in FIG. 10.

A possible explanation for this can be traced back to the fact that the city of Brindisi is the location of Italy's largest coal-fired power plant, which has been a serious national environmental problem for some time by now[6][7].

Looking at the data in the TAB. V we noticed that the maximum values of PM2.5 are found on days with strong winds. Given this consideration, it is possible that the wind carried the polluting particles from ILVA in Taranto[8]. We know, in fact, that as the largest steel plant in Europe, it greatly pollutes the surrounding environment and probably the wind helps the dispersion of the polluting particles. In fact, the two cities are only 60 km apart as the crow flies.

Furthermore, confirming what was said in the previous sections, it can be seen that all 5 observations are in the coldest months, characterised by a known worse air quality.

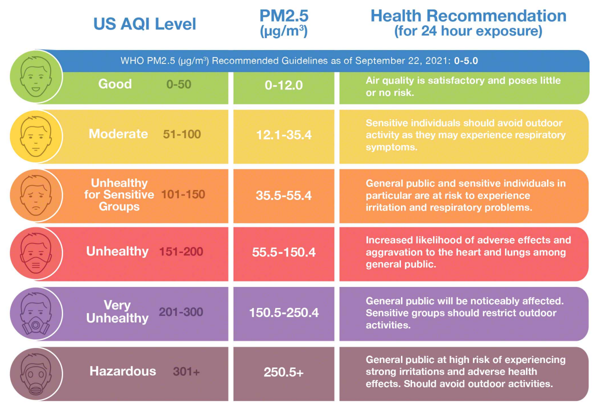

FIG. 10: Air Quality Index.
Source: IQAir and U.S. EPA.

CONCLUSIONS

During the development of this project, a final database was obtained containing pollution parameters and weather conditions for days in the year 2022 for a selection of European states. This allowed us to perform exploratory analyses on the relationship between weather and air quality.

Limitations

There were a few limitations that hampered the initial aims of the analysis. One of these was the absence of the pollutant CO2 among the parameters obtained from the first API. We know, in fact, that carbon dioxide is one of the primary causes of rising global temperatures and one of the most relevant pollutants today. Furthermore, with regard to the analysis centred on Italy, the first API unfortunately does not return data for the Lombardy region, which is known to register the highest national pollution levels.

Future developments

By answering the last two research questions focused on Italy, we observed that the trends in pollutants appeared more evident and were more in line with the theoretical considerations made initially. A possible future development of this project could see an in-depth analysis for each state taken individually, rather than condensing states into geographical regions that might contain states with different climatic conditions.

References

1. IQAir - How wind and weather affect air pollution. https://www.iqair.com/us/newsroom/wind-weather-air-pollution.
2. IQAir - Does air pollution cause climate change? https://www.iqair.com/us/newsroom/does-air-pollution-cause-climate-change.
3. IQAir - Air quality in the world. https://www.iqair.com/us/world-air-quality.
4. National Oceanic and Atmospheric Administration. https://www.ncei.noaa.gov/.
5. United Nations Geoscheme for Europe. https://en.wikipedia.org/wiki/United_Nations_geoscheme_for_Europe.
6. Brindisi, la centrale più inquinata d'Italia. https://www.wwf.it/pandanews/ambiente/brindisi-la-centrale-piu-inquinante-ditalia/.
7. Centrale termoelettrica Federico II di Brindisi. https://www.wwf.it/area-stampa/centrale-termoelettrica-federico-ii-di-brindisi/.
8. Inquinamento atmosferico causato dall'ILVA. https://www.controlsecurityambiente.com/inquinamento-atmosferico-causato-dallilva/.