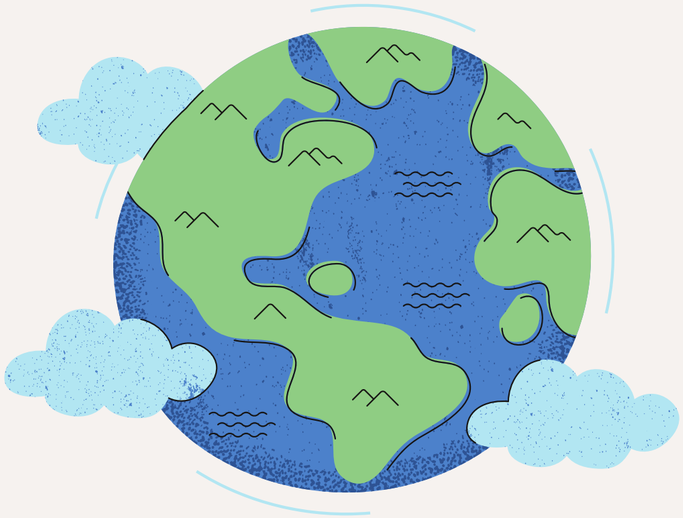
Exploring the relationship between pollution and weather conditions: an European analysis



Andrea D’Amicis, Gaia Gubelli, Emmanuele Lotano

Data Management project
Academic year 2023/2024

Main goal

This study aims to provide a descriptive overview of the dynamics governing **air pollution** in relation to **weather conditions**.

The goal is to obtain a unique database, which contains both metereological conditions and pollution measurements.

The data collected refer to the **year 2022** and encompass the majority of **European States**.

Research questions

European descriptive analysis:

1. What are the European average O3 levels across maximum temperature ranges?
2. What are the European average NO2 levels in the precipitation ranges?
3. What are the European average PM10 levels in the different wind speed ranges?
4. What are the European average values of PM10 by season?

Italian descriptive analysis:

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

European states included

- 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)

Data acquisition

Pollution

- Retrieval of the data from the API: https://openaq.org
- For each city available we obtained the daily measurements averaging the hourly detections

The parameters requested in each call are:

- **PM10**: particulate matter with diameter ≤ 10 *μm*
- **PM2.5**: particulate matter with diameter ≤ 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 *μg/m3*.

Weather

- Retrieval of the data from the API: https://open-meteo.com
- For each city of the first dataset, we obtained the daily meteorological measurements

The parameters requested in each call are:

- **WMO code**
- **Max. temperature** (°*C*)
- **Min. temperature** (°*C*)
- **Wind Speed** (*km/h*)
- **Apparent max. temperature** (°*C*)
- **Apparent min. temperature** (°*C*)
- **Precipitation** (sum in *mm*)
- **Rain** (sum in *mm*)
- **Snow** (sum in *mm*)

Storage

- **Document-oriented** database: MongoDB
- Documents structured in a BSON format
- Fields can have sub-documents
- Schema-less which gives greater **flexibility** and **efficiency** when dealing with large datasets
- At the end of this phase we obtained **180’768** documents for each collection.


```
_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
```

```
_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
```


Examples of documents in the two raw collections

Data profiling: completeness

Pollution

Here, zeros are actually missing values.

Table completeness:
- 35% (8% + 27%) of missing values

Attribute completeness:
- missing values in “NOx”, “CO” and “NO” represents almost the totality of the dataset

	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.

Weather

Here, zeros represent correct measurements, indicating the absent of a certain phenomenon.

Table completeness:
- there are 45'183 missing values, <1% of the entire dataset

Attribute completeness:
- number of missing values is minimal

	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.

Data profiling: consistency

Pollution

The idea is to assess the consistency of the values with reference to real observable values.

- There are **negative** values, which are impossible
- Some maximum values are improbable

--→ probably due to measurement errors or non-natural phenomena (e.g. 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

Looking at the ranges we can observe some inconsistencies:

- **WindSpeed** and **PrecipitationSum** show negative values
- Maximum value of **TemperatureMin** is higher than the maximum value of **TemperatureMax**.

--→ there are some documents whose measurements are completely inconsistent

	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.

Data Integration and Cleaning

We merged the two datasets to obtain for each day and for each city, a single coherent view of parameters.

Pollution

- Removal of documents with **excessively high** values of pollution parameters
- Removal of documents with **negative values** of pollutants
- Removal of **CO**, **NOx** and **NO**, which contained almost all NaN
- **Mean replacement** for remaining NaN, for each state taken individually

Weather

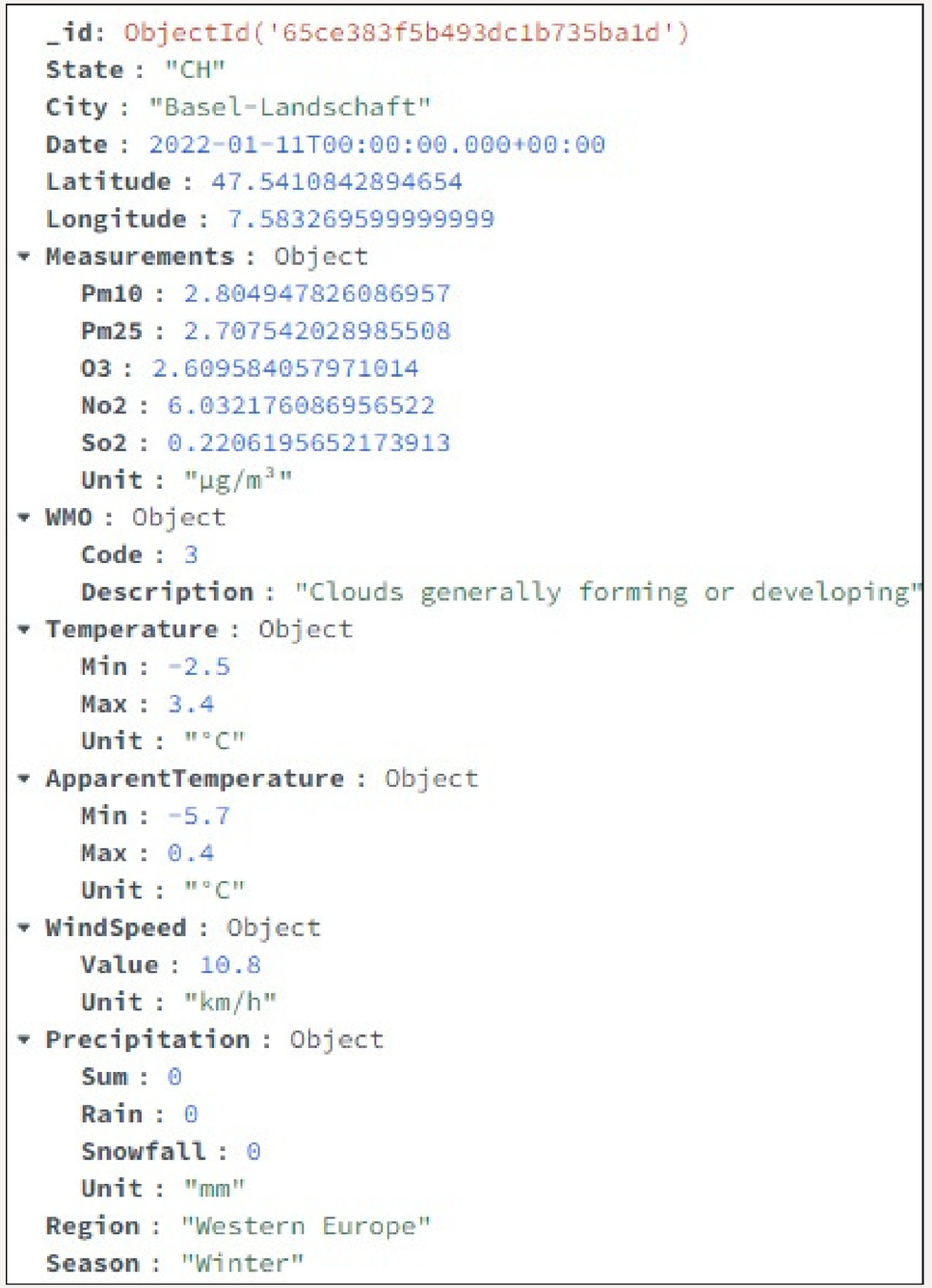
- **Inconsistent** and **improbable** measurements. **Removal** of documents where:
  - Min. temperature > Max. temperature
  - Apparent min. temperature > Apparent max. temperature
  - Min. temperature > 50 OR Apparent min. temperature >0
- Documents with inconsistent values of **WindSpeed** and **Precipitation** were the same of those with inconsistent temperature values

Data Enrichment

- Addition of the **description related to WMO_code** to provide greater interpretability:
  - for example, WMO = 1 is “Cloud development not observed or not observable”
- Addition of **Season** attribute (Winter, Summer, Spring, Autumn)
- Addition of **Region**, according to the United Nations Geoscheme for Europe:
  - *Northern Europe:* Norway, Finland, Sweden, Great Britain and Ireland
  - *Western Europe:* Germany, Austria, Switzerland, Netherlands, Belgium and France
  - *Southern Europe:* Italy, Spain and Portugal

Final Storage

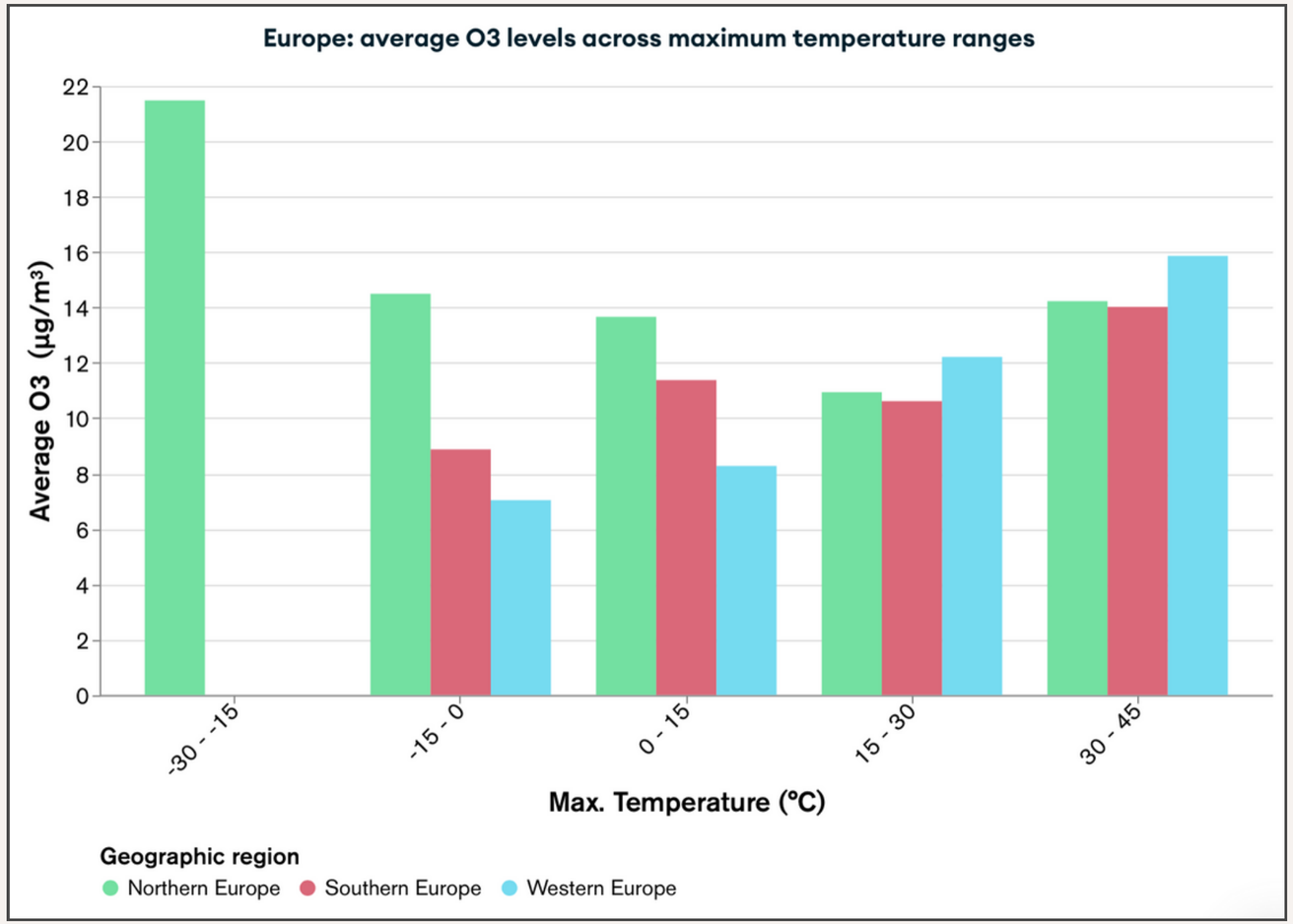
- We obtained a collection containing 99'048 documents
- **Measurement** object contains pollution parameters
- **Weather** parameters are clearly defined
- Units of measurement are provided for each parameter



Queries

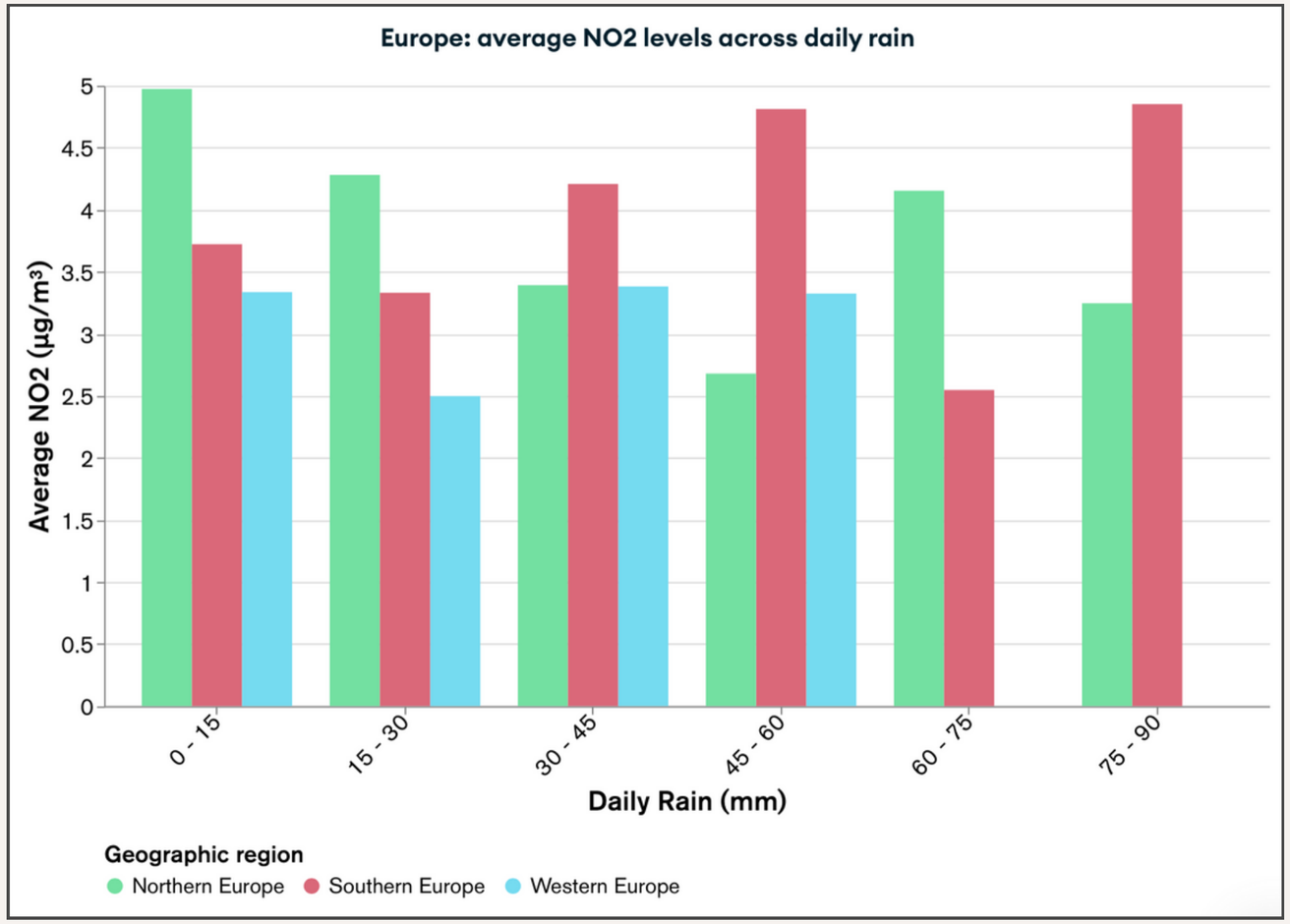
European average O3 levels over temperature

- The trend in O3 levels seem to reflect the theoretical considerations with the exception of the **Northern Europe** region
- O3 levels are higher for warmer temperatures
- Temperature range [-30, -15] stands out: it encompasses obs. from Scandinavian countries



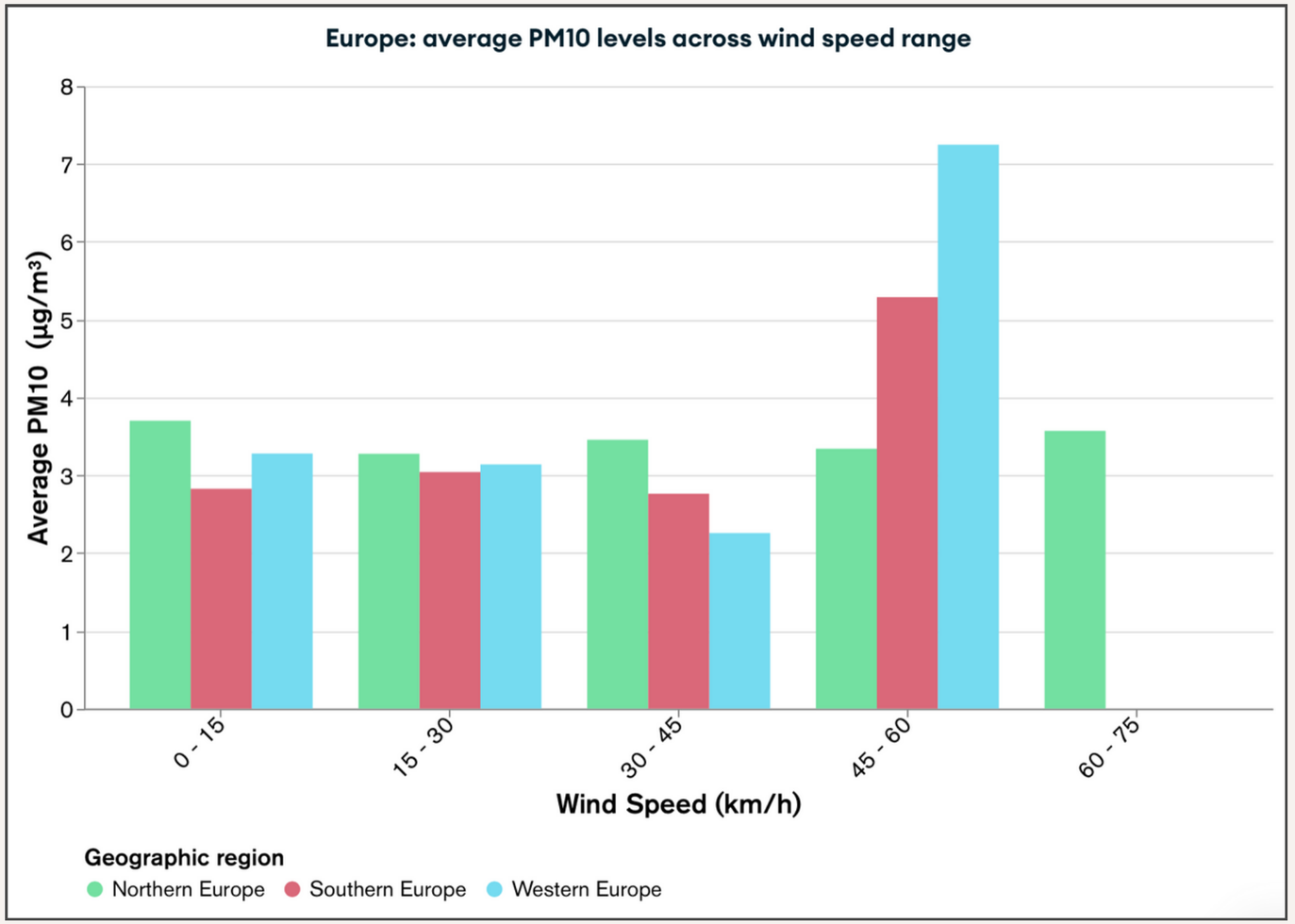
European average NO2 levels over daily rain

- There is no clearly defined trend
- The phenomenon represented is complex:
  - cleaner air is probably observed the day after a rainy day
- Classes [90,105] and [105+] were removed, as they were represented by only 2 documents each



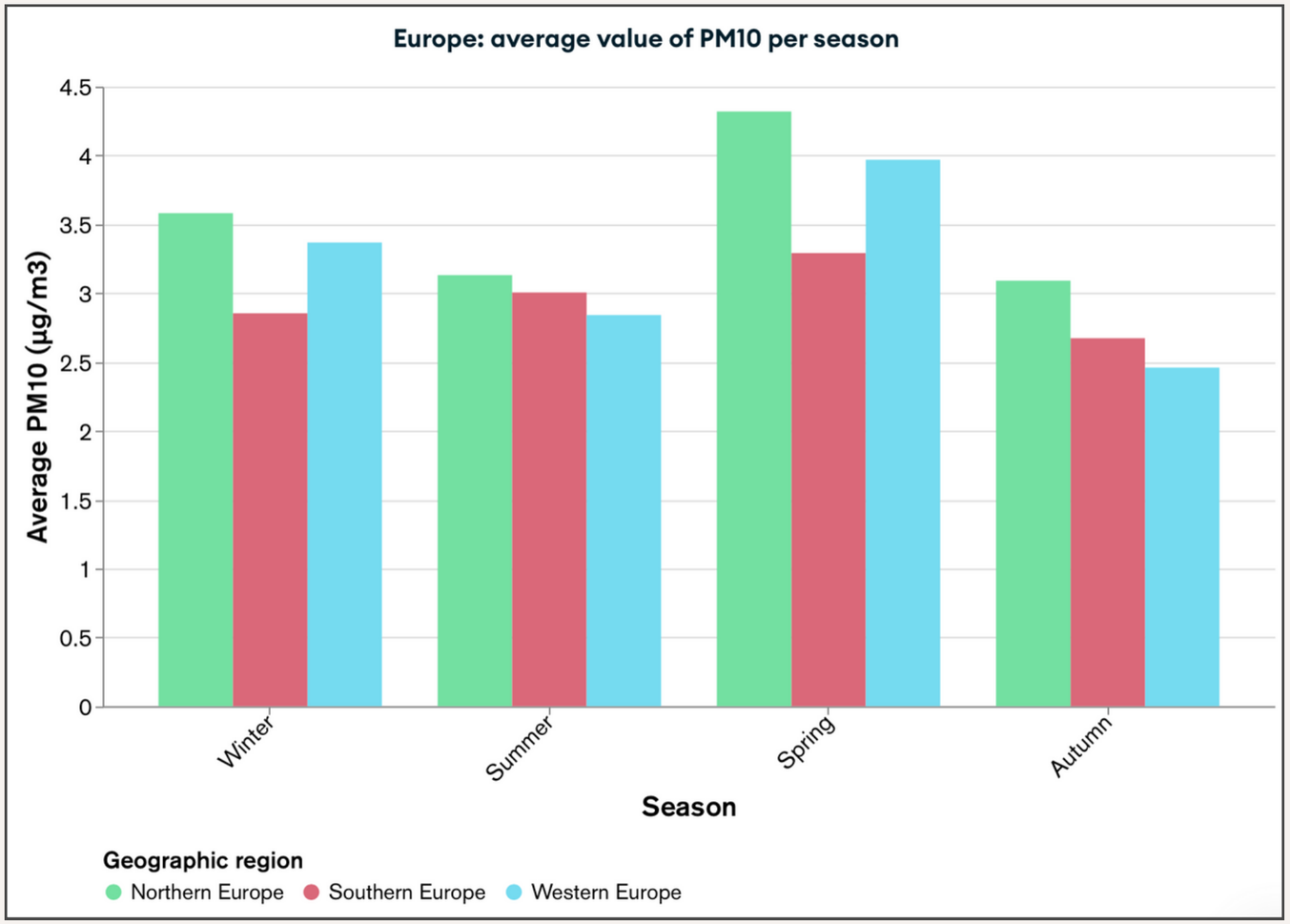
European average PM10 levels over wind speed

- First three classes are homogenous
- Class [45,60]: clear difference in the increase of the average PM10 value (particularly for Western Europe)
- Class [75,90] was represented by only 2 documents and therefore it was excluded from the analysis



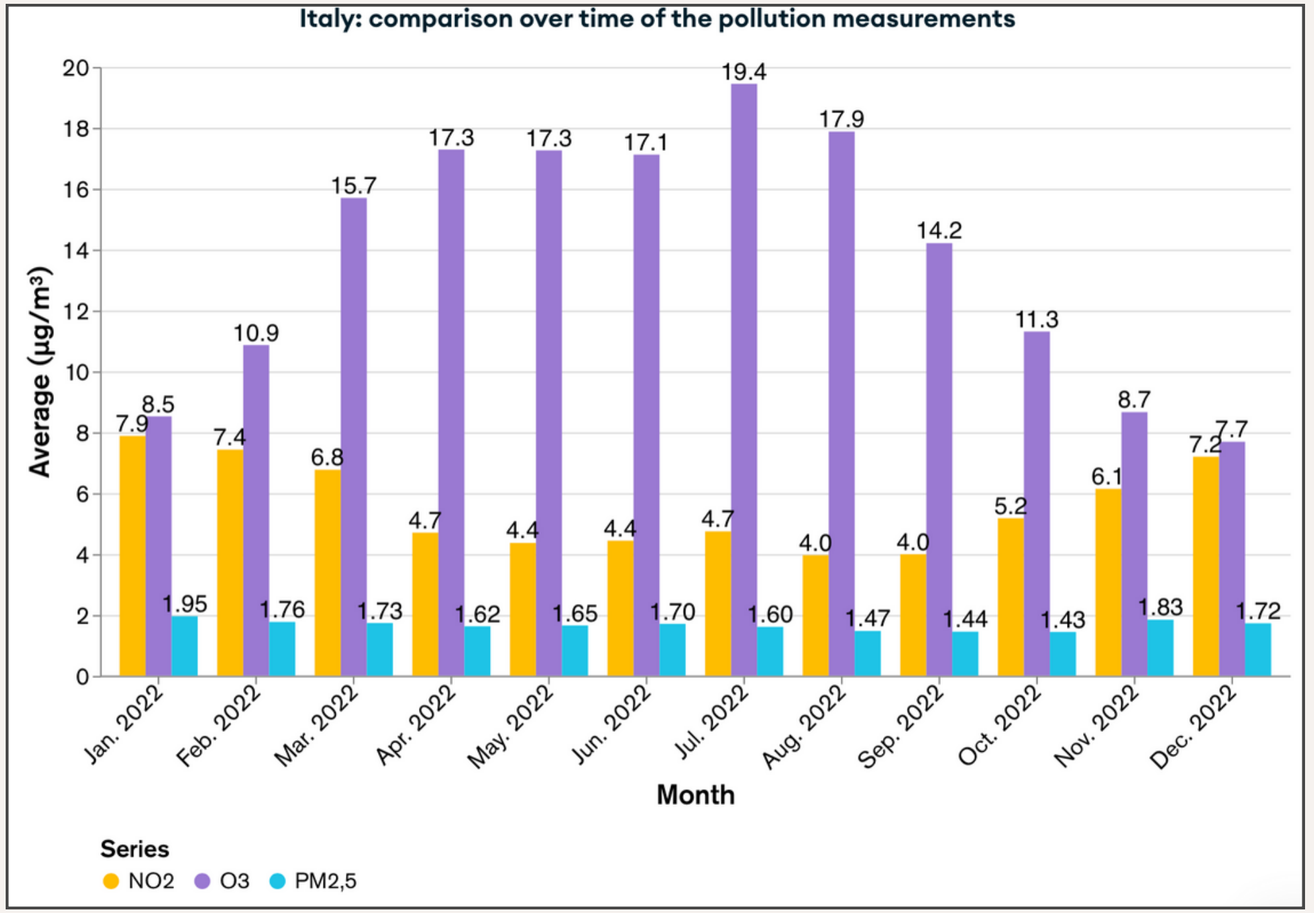
European average PM10 across seasons

- Highest average levels observed in the spring:
  - PM10 is mainly composed of dust, pollen and spores
- Winter also shows slightly higher values (with the exception of the Southern Europe regions)



Italian pollution measures over time

- **NO2**:
  - compound composed of gaseous polluting particles produced by industrial processes and domestic heating
  - it shows higher values during colder months
  - it shows lower values during the warmer months
- **O3**:
  - opposite trend
  - maximum levels are reached in the summer
  - Ozone is formed on the days with higher temperature
- **PM2.5**:
  - average levels remain constant throughout the year
  - slightly higher values occur in the colder months



Highest PM2.5 values observed in Italy

- The first five documents with the highest PM2.5 values are observed in **Brindisi**
- According to the Air Quality Index, these values are considered a problem for health
- **Hypothesis** of the cause:
  - in Brindisi there is the Italy's largest coal-fired power plant
  - only 60 km apart as the crow flies there is the steel plant ILVA, in Taranto
- These high values are observed in **strong wind days** and in the **coldest months**

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.

Conclusions

Limitations:

- absence of CO2 parameters from the API
- the pollution API did not made available cities of Lombardy

Future developments:

- in-depth analysis for each state taken individually

Thanks for the attention