Air Quality in Italy

Enrico Paramucchi

February 19, 2020

1. Introduction

1.1 Background

Air quality is deteriorating over the last century, becoming nowadays a problem for public health. It is globally recognized that air quality conditions have a direct impact on the heath of citizens.
As such, air samples are periodically analyzed by Authorities worldwide.
In Italy this activity of monitoring concentrations of pollutants in air is performed by Regional Environmental Agencies that submit their data to the European Environment Agency (EEA).
And indeed, the EEA database is the source of data analyzed in this report.
The overall purpose is to group Italy's territory in homogeneous areas characterized by similar air quality conditions in order to understand and possibly predict the impacts of air pollutants on local citizen.
The air pollutant considered is Nitrogen (oxides, monoxides and dioxide), amongst the most dangerous for human health, as citizens living in areas with high concentrations of nitrogens are more likely to become sick due to poor air quality.

1.2 Problem

Data that might contribute to define homogenous areas based on concentrations of nitrogens should include location info (longitude and latitude), pollutants and their concentrations.
This project aims to cluster Italy neighborhoods to define similar areas based on the current average levels of pollutants to anticipate health issues pattern throughout the country.

1.3 Interest

Environment, Health and Local Authorities may be interested in the finding, as similar clusters may be talked in similar ways.
At the same time citizens may have a skate, as they may use the finding of this piece of research to identify the better areas to live in Italy (in terms of air quality) or to consider mitigation actions (e.g. installation of air purifier systems in their houses).

2. Data acquisition and cleaning

2.1 Data sources

As stated above the data source is the European Environment Agency database, where concentrations of *nitrogens* have been retrieved as follows:

- Italy territory
- Annual mean for 2018
- Nitrogen oxides, monoxides and dioxide

Link to the query is here provided.

2.2 Data cleaning

Data downloaded in csv were converted into a python dataframe.

	CountryOrTerritory	ReportingYear	UpdateTime	StationLocalId	SamplingPointLocalId	SamplingPoint_Latitude	SamplingPoint_Longitude	Pollutant	AggregationType
0	Italy	2018	2019-10-29T22:23:39.82Z	STA.IT1159A	SPO.IT1159A_38_chemi_1998-01-02_00:00:00	44.499720	11.328330	Nitrogen monoxide (air)	Annual mean / 1 calendar year
1	Italy	2018	2019-10-29T22:23:39.82Z	STA.IT1728A	SPO.IT1728A_38_chemi_2004-03-01_00:00:00	42.552220	12.651670	Nitrogen monoxide (air)	Annual mean / 1 calendar year
2	Italy	2018	2019-10-29T22:23:39.82Z	STA.IT2252A	SPO.IT2252A_38_chemi_2016-05-18_00:00:00	42.081825	11.809336	Nitrogen monoxide (air)	Annual mean / 1 calendar year
3	Italy	2018	2019-10-29T22:23:39.82Z	STA.IT2064A	SPO.IT2064A_38_chemi_2009-06-20_00:00:00	43.273889	12.611667	Nitrogen monoxide (air)	Annual mean / 1 calendar year

Of the many fields available, only the following ones have been retained:

- StationLocalId
- SamplingPoint_Latitude
- SamplingPoint_Longitude
- Pollutant
- AQValue

With the following meanings:

- Unique Identifier of the Sampling Point
- The Latitude of the Sampling Point
- The Longitude of the Sampling Point
- The type of Pollutant that has been measured
- The Concentration on the pollutant (in ug.m-3)

Features are then selected as schematized below:

Field ID	Technical Name	Description
1	StationLocalId	Unique Identifier of the Sampling Point
2	SamplingPoint_Latitude	The Latitude of the Sampling Point
3	SamplingPoint_Longitude	The Longitude of the Sampling Point
4	Pollutant	The type of Pollutant that has been measured
5	AQValue	The Concentration on the pollutant (in ug.m-3)

The dataframe is then split into two:

- A first one listing sample points and related Longitude & Latitude
- A second one listing the Pollutants and related Concentrations

Duplicated values are removed from the first dataframe, in order to keep only one unique value by StationLocalId.

In the second one, Pollutants are firstly moved from rows to columns, then rows with NaN in terms of values (missing concentration values) are deleted:

From ROWS

	StationLocalId	Pollutant	AQValue
0	STA.IT1159A	Nitrogen monoxide (air)	22.998201
1	STA.IT1728A	Nitrogen monoxide (air)	6.416513
2	STA.IT2252A	Nitrogen monoxide (air)	2.512227
3	STA.IT2064A	Nitrogen monoxide (air)	4.297892
4	STA.IT1152A	Nitrogen monoxide (air)	13.022490
5	STA.IT1590A	Nitrogen monoxide (air)	16.960381
6	STA.IT1596A	Nitrogen monoxide (air)	6.368159
7	STA.IT0861A	Nitrogen monoxide (air)	51.916251
8	STA.IT1273A	Nitrogen monoxide (air)	0.501961
9	STA.IT2203A	Nitrogen monoxide (air)	1.616236

To COLUMNS

Pollutant	Nitrogen dioxide (air)	Nitrogen monoxide (air)	Nitrogen oxides (air)
StationLocalId			
STA.IT0063A	14.939116	2.973279	19.491825
STA.IT0187A	37.588794	23.639182	72.853887
STA.IT0267A	24.291143	NaN	NaN
STA.IT0448A	28.281029	14.652279	49.823091
STA.IT0459A	29.447774	NaN	NaN
STA.IT0460A	27.610241	NaN	NaN
STA.IT0461A	27.202162	NaN	NaN
STA.IT0469A	51.899691	NaN	NaN
STA.IT0470A	56.016281	NaN	NaN
STA.IT0477A	59.259929	NaN	NaN

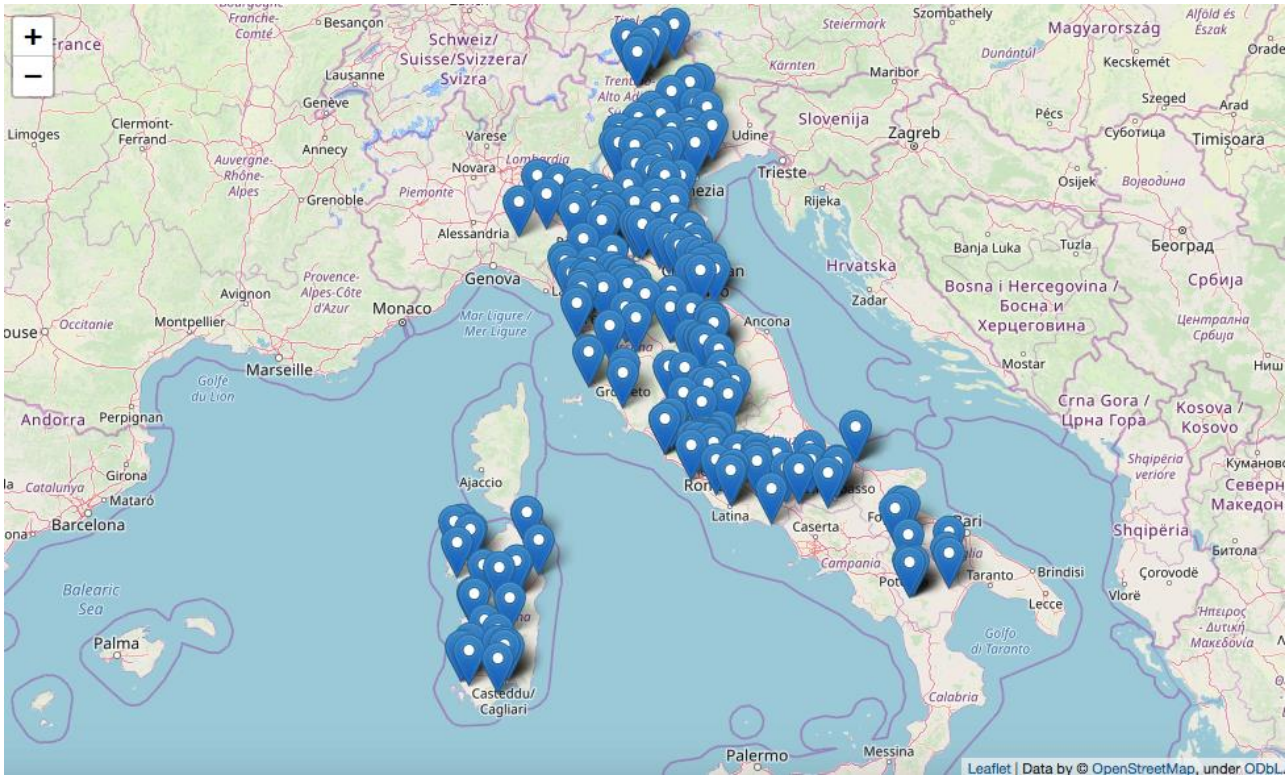
The two clean sub-dataframes are finally re-merged into one final dataframe listing all sampling points and related spatial coordinates where pollutants concentrations means have been measured in the course of 2018:

	StationLocalId	Nitrogen dioxide (air)	Nitrogen monoxide (air)	Nitrogen oxides (air)	SamplingPoint_Latitude	SamplingPoint_Longitude
0	STA.IT0063A	14.939116	2.973279	19.491825	42.939167	10.534167
1	STA.IT0187A	37.588794	23.639182	72.853887	44.842500	11.613060
2	STA.IT0448A	28.281029	14.652279	49.823091	45.429444	12.313889
3	STA.IT0502A	39.826836	33.981320	91.930023	46.482310	11.341830
4	STA.IT0505A	3.066276	0.253706	3.461377	46.589670	11.433170

3. Exploratory Data Analysis

3.1 Sampling Points

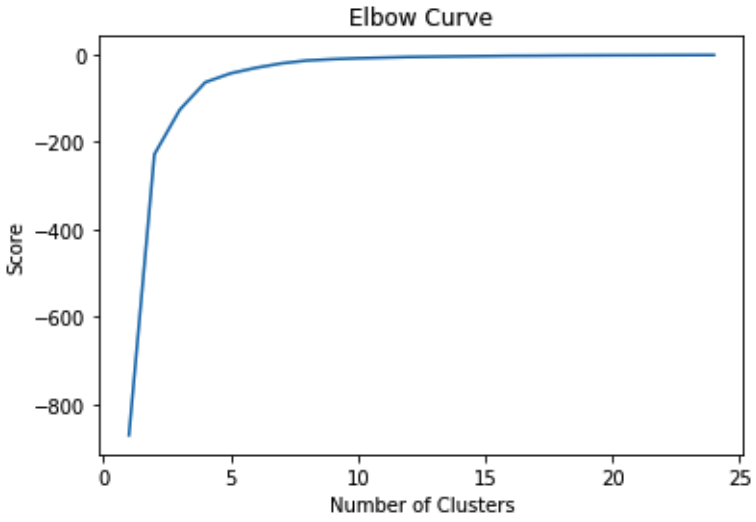
The position of each Sampling Point is presented on a map:



It is immediately clear that Italy is not covered throughout its territory. Indeed, coverage is not secured for the north-west part as well as the south. This means that these territories cannot fall within the analysis, by reducing the overall scope

3.2 Elbow Curve

An elbow curve based on spatial data is then elaborated in order to identify the optimal number of clusters, that results to be 4 as presented by the graph below:



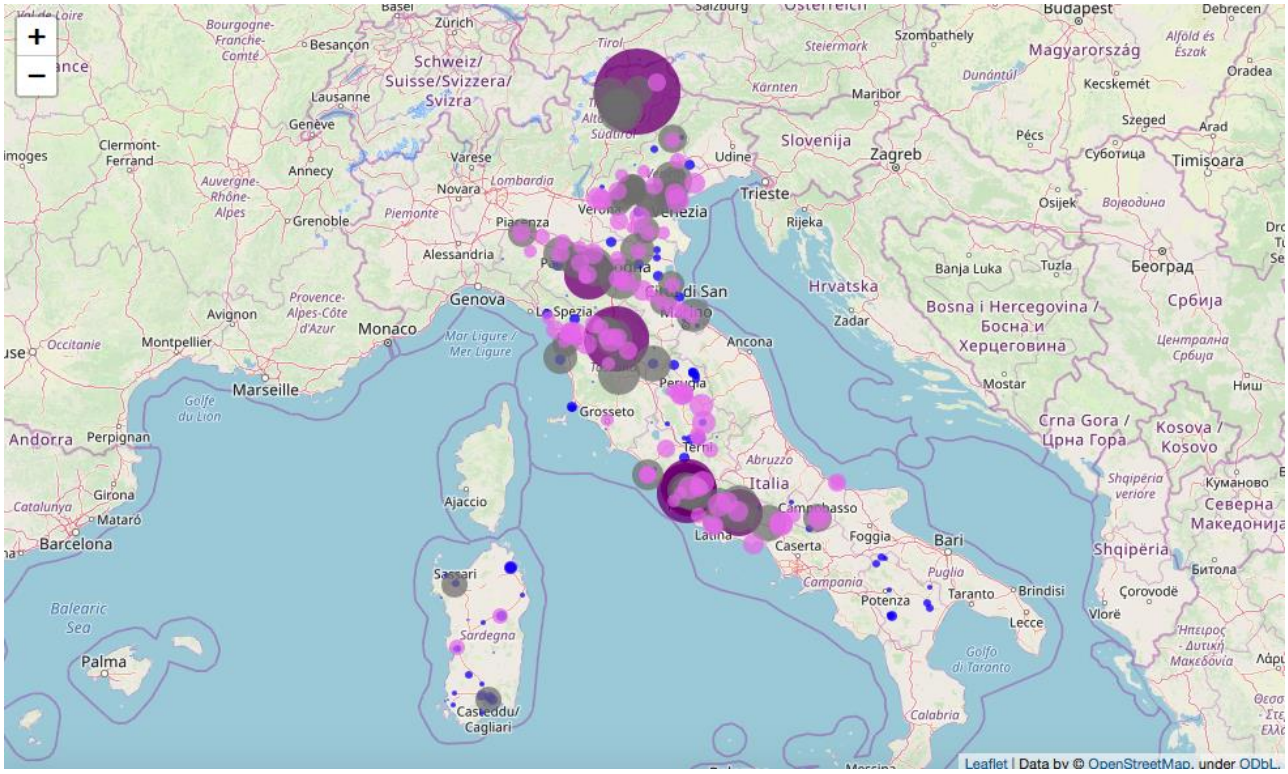
3.3 K-Means

Based on the “elbow curve” a K-Means iteration is started and completed with the clustering of the dataframe:

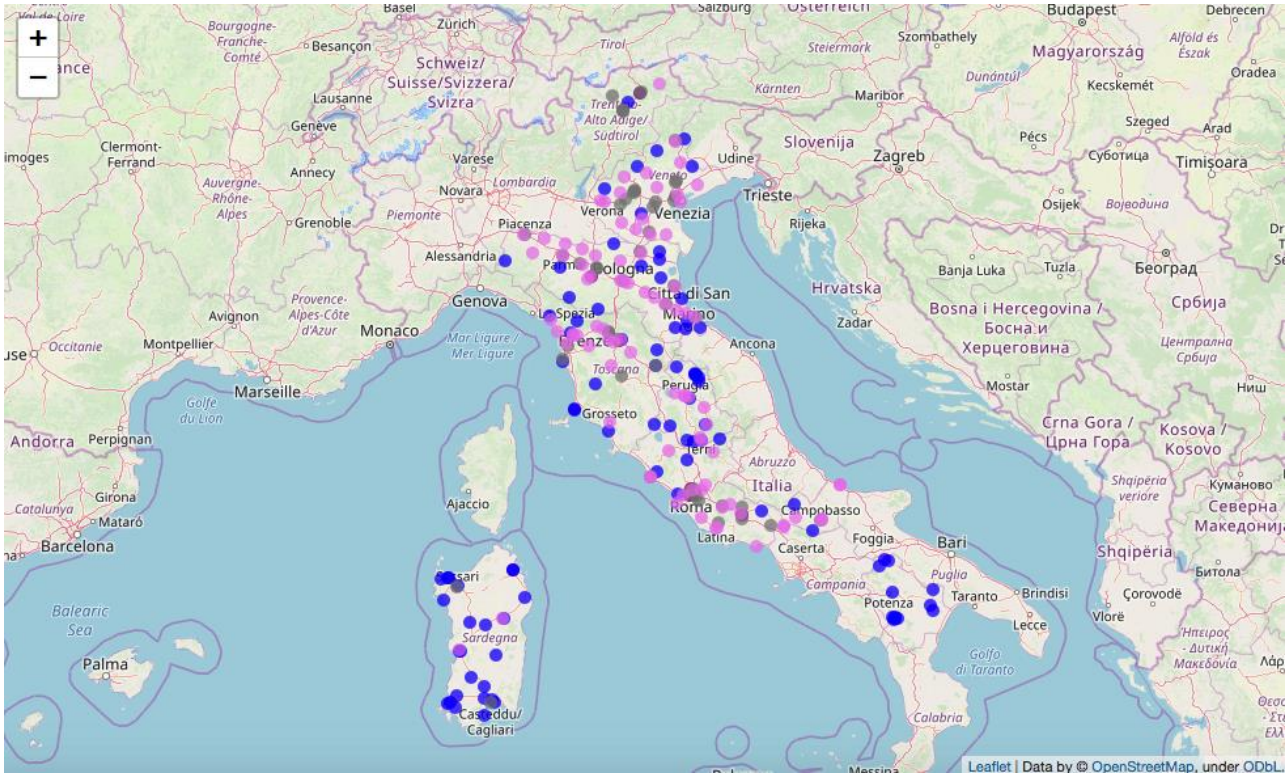
	SamplingPoint_Latitude	SamplingPoint_Longitude	Total_Nitrogens	Cluster_Label
0	42.939167	10.534167	37.0	0
1	44.842500	11.613060	134.0	2
2	45.429444	12.313889	93.0	3
3	46.482310	11.341830	166.0	2
4	46.589670	11.433170	7.0	0
5	46.714920	11.654080	114.0	2
6	46.797340	11.944030	71.0	3
7	44.823890	9.830280	52.0	3
8	45.714444	11.368611	54.0	3
9	44.636050	10.904730	147.0	2

The corresponding “spatial results” are as follows:

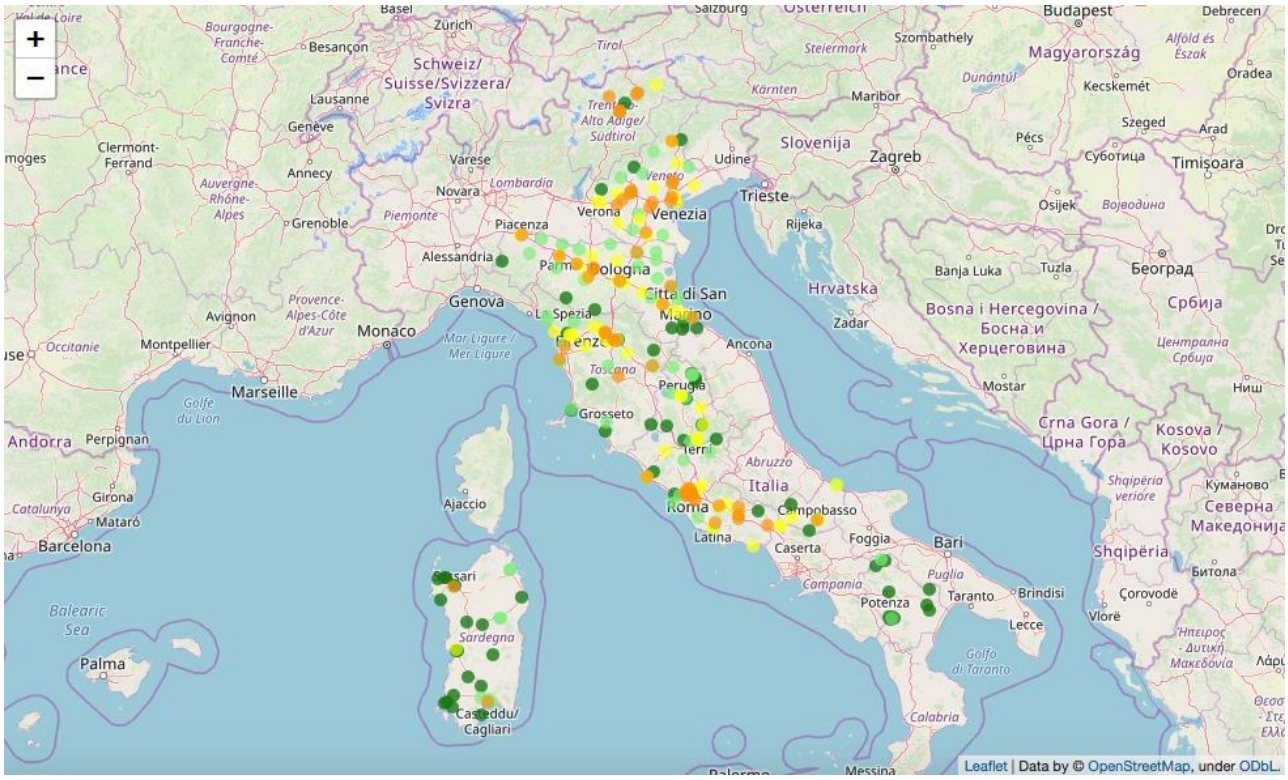
- From the first map we assess the assumption that clusters of the same color should have approximately the same dimension, that appears to be true.
- This means that clustering has been performed by Concentration Levels, as the dimension of cluster is given by the formula: radius = 0.1 * nitrogens



The second map here below presents the same information of the one above, but without dimensioning the cluster based on the related concentrations of nitrogens:



This second one can be compared against the third map below, where sample points are colored not by cluster but by concentration of nitrogens from light-green (low concentrations) to orange (high concentrations):



Although not perfectly overlapping with clusters, the binning of the dataframe into quartiles confirms the same pattern, i.e. areas close to big centres (e.g. Rome, Florence and Bologna) are the most polluted, together with the north-east side of the country and the Pianura Padana.

5. Conclusions

Examination of air pollutants according to sampling points throughout Italy shows that

1. North East Italy as well as South Italy have been excluded due to lack of sample data;
2. Areas close to big centres (e.g. Rome, Florence and Bologna) are the most polluted, together with the north-east side of the country, the densest in terms of factories and the Pianura Padana.

Examination of pollutants by clusters shows:

1. Clusters 0 and 1 identify the ‘less polluted’ areas;
2. Clusters 2 and 3 are characterized by higher concentrations of air pollutants.

For the case just analyzed, K-means clustering doesn’t really provide added value compared to a more basic categorization of sampling point quartiles based on the concentration of pollutants detected.

This means that the model needs to be further refined in order to include additional drivers such as number of inhabitants and related age.

This can determine the identification of homogeneous clusters to be used to address health issues: clusters where concentration of pollutants is high, the number of inhabitants is high and the average age is high as well should be tacked differently from a cluster where concentration of pollutants is still high, but the number of inhabitants and related aging is lower.