Project proposal of Visual Analytics

Elisa De Bellis
1858927

Dataset

The dataset is taken from ISPRA (Istituto Superiore per la Protezione e la Ricerca Ambientale), which provides a comprehensive overview of air quality in various regions of Italy, with detailed data on air pollutants collected from monitoring stations distributed throughout the territory. The pollutants taken into consideration:

- PM2.5 and PM10 are particulate matter particles (depending on the size 10μm and 2.5μm) in the air, produced by natural causes (forest fires, soil erosion processes, etc.) that cause serious respiratory problems.
- Nitrogen dioxide (NO2) is a pollutant emitted mainly by vehicle traffic; other sources include civil and industrial heating plants and power plants. Nitrogen dioxide has negative effects on human health and, together with nitrogen monoxide, contributes to smog.
- Ozone (O3) is the main smog pollutant. The highest ozone concentrations occur during the hottest months of the year and during hours of maximum solar radiation. The main sources of ozone precursor compounds are road transport, domestic heating and energy production. Ozone can cause serious problems to human health and the ecosystem, as well as to agriculture and material goods.

To establish air quality on the basis of previous pollutants, reference is made to the European Environment Agency (EEA), which provides the following guidelines:

Pollutant	Index level (based on pollutant concentrations in µg/m3)					
	Good	Fair	Moderate	Poor	Very poor	Extremely poor
Particles less than 2.5 µm ($PM_{2.5}$)	0-10	10-20	20-25	25-50	50-75	75-800
Particles less than 10 µm (PM_{10})	0-20	20-40	40-50	50-100	100-150	150-1200
Nitrogen dioxide (NO_2)	0-40	40-90	90-120	120-230	230-340	340-1000
Ozone (O_3)	0-50	50-100	100-130	130-240	240-380	380-800

Figure 1: European Air Quality Index

General idea

The objective is to analyse air quality data in order to understand trends in air pollutants and identify areas with the highest pollution. The analysis will also focus on identifying the most critical periods and regions. The user can interact with the map to see the different pollutant concentrations and the general situation in the various regions and cities. In addition, the user can view the corresponding values via a box plot with reference to the permitted European average threshold. It is also possible to view developments over time.

Intended user

Intended users include environmental researchers, environmental policy makers, government agencies and non-governmental organisations (NGOs) dealing with air quality and public health. The visual analytics tool will help support data-driven decision-making, improve public awareness and assess the impact of environmental policies.

Visual Analytics cycle

The main parts of Visual Analytics cycle:

Analytics

As a preprocessing part, missing data will be handled, normalization of the data will be done to ensure that all characteristics contribute equally to the analysis, especially if different pollutants have very different scales. In addition, some aggregations on days/months will be made to analyse trends over various periods. I think to use t-SNE as the dimensionality reduction in order to obtain clusters and similarities between different regions or time periods based on pollutant concentrations. Or maybe even PCA to capture the maximum variance in the data.

Visualization

Visual part is constituted by 4 visualizations:

- Interactive Map: which shows a map of Italy with values corresponding to the air quality in each geographical area indicated.
- Box-Plot: showing the data against the permitted European threshold
- Scatter-Plot: showing the result of dimensionality reduction
- Time-Series plot: showing the evolution of data over time

Interaction

The user will be able to interact with the map by selecting a particular region and city, waiting for a change in the other graphics in real time corresponding to a subset of data. Similarly, when a particular year is selected in the time-series plot, the data for that year will be shown in the other graphs. In addition, the user will be able to select certain clusters from the scatter plot and see the corresponding data on the other graphs.

The proposal is just a draft of the project, as it unfolds I might want to change a few things.

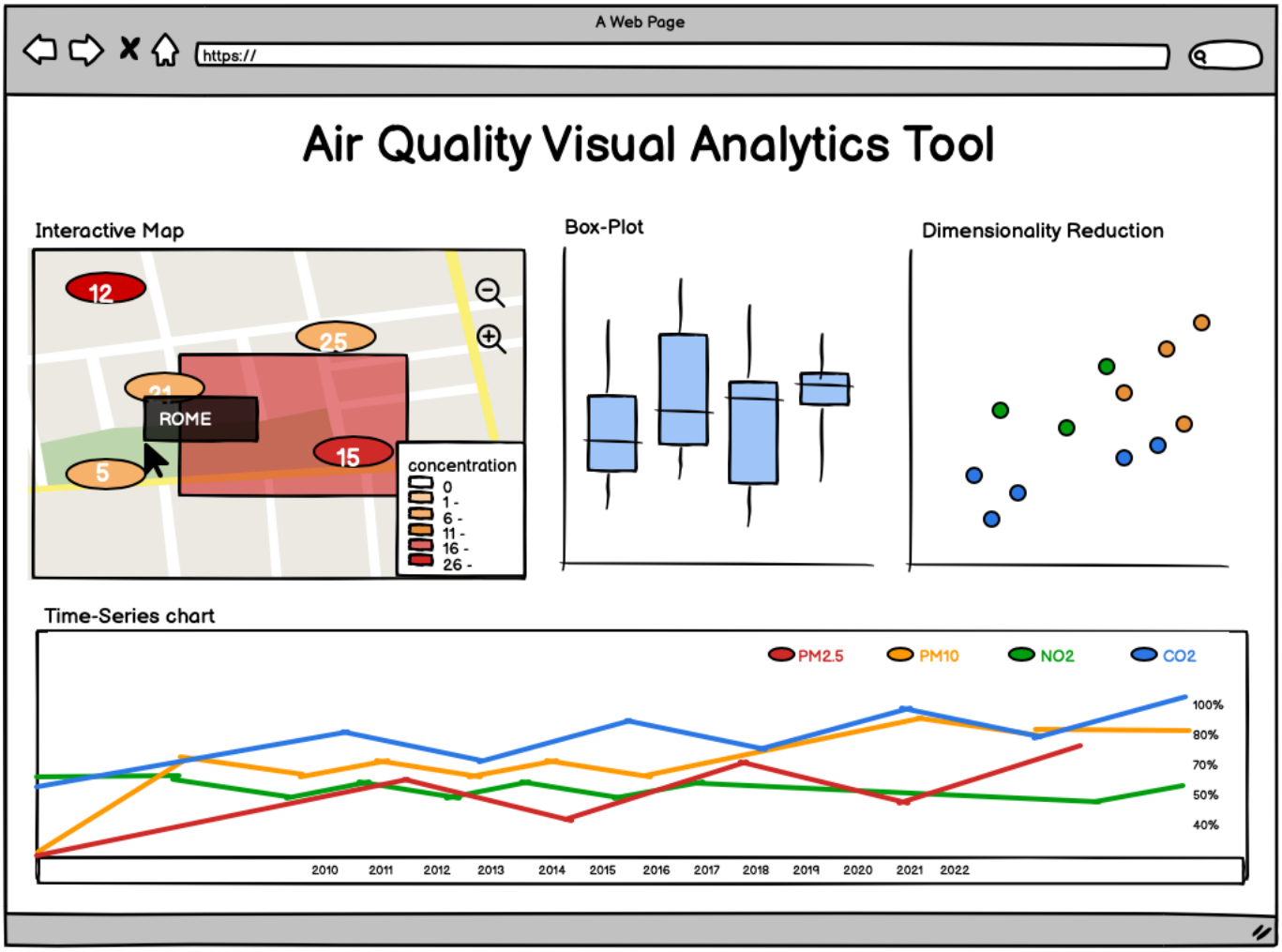

Figure 2: Draft mock-up of the user interface