VISUAL-ANALYTICS TOOL FOR AIR QUALITY INDEX

Elisa De Bellis 1858927

Sapienza Univeristy of Rome
<debellis.1858927@studenti.uniroma1.it>

ABSTRACT

Air pollution emissions have declined in the last decade, resulting in better air quality. Despite this improvement, air pollution remains the largest environmental health risk in Europe and Italy. This paper presents Air Quality Index Visual-Analytics tool to analyze concentration of pollutants in Italy in order to improve public awareness and assess the impact of environmental policies. The tool offers an interactive map, time-series, scatter-plot and box-plot to identify patterns. Users can interact with the tool to explore insights such as region with higher concentration of some pollutant, or to capture an overview of the last 10 years about the situation of Italy.

Keywords: *Air quality index, Pollutants, Italy, Visual-Analytics*

1. INTRODUCTION

Air pollution is currently one of the greatest environmental threats to human health in Italy, Europe and the world. Although air quality is improving and the emission of all major air pollutants in the European Union is decreasing over time, according to many experts, pollution should be perceived as the second biggest environmental threat after global warming. As of 2019, Italy ranks first in the European Union for premature deaths due to pollution. Some 63,700 pollution-related premature deaths were recorded in Italy in 2019.

The project aims to analyze the air pollution emissions from 2010 to 2021, and identify and classify which are the region with the worst situation. This aims to develop an analysis and a tool that can be used both at the planning level to define appropriate measures for reducing emissions, and at the operational level, to promptly intervene in areas with excessively high levels of pollutants. The whole project was developed using React, JavaScript and the D3.js library.

2. RELATED WORKS

In recent years, the issue of air quality has received significant attention globally due to its impact on public health, environmental sustainability, and socioeconomic factors. In this section, relevant tools in the literature are reviewed focusing on key themes within air quality and pollutant emissions studies.

1. The TRAFAIR air quality dashboard (BACHECHI et al., 2020), a web application designed to monitor and visualize urban air quality conditions. This tool aims to help decision-makers understand air quality patterns and implement effective policies. The dashboard collects data from a network of low-cost sensors installed in various urban locations. These sensors measure concentrations of key pollutants, such as carbon monoxide (CO), nitrogen dioxide (NO2), nitrogen monoxide (NO), and ozone (O3), providing real-time data every two minutes. This spatio-temporal data is then visualized using dynamic and interactive graphics, which display real-time statistics and trends of air quality conditions across the city.

2. AirVIS (LIAO et al., 2014), a web-based system designed for comprehensive analysis of air quality data. The system integrates visual analytics techniques to process and analyze spatial-temporal and multi-dimensional air quality data. It provides three primary visual views:

   - GIS View: Displays the spatial distribution of air quality data using a map with pie charts showing pollution levels.
   - Scatter Plot View: Visualizes temporal patterns of air quality data over time.
   - Parallel Coordinates View: Analyzes correlations between multiple pollutants and their impact on the Air Quality Index (AQI).

The system aims to help users understand the distribution and trends of air pollutants, identify anomalies, and explore the relationships between different pollutants.

3. Air Quality Index (AQI) (EEA,) The AQI is one of the most widely used tools globally to monitor air quality in real time. It provides a numerical value based on the concentration of pollutants such as PM10, PM2.5, NO2, SO2, CO and O3. The index varies from 'good' to 'dangerous' depending on pollution levels, allowing the public to easily understand the state of the air. It is used by bodies such as the US Environmental Protection Agency (EPA) and the European Environment Agency (EEA) European Air Quality Portal. The air quality portal of the European Environment Agency (EEA) collects and publishes data from thousands of monitoring stations across Europe. Data includes levels of particulate matter (PM10 and PM2.5), nitrogen oxides (NOx), ozone (O3) and other pollutants. It is used to provide analysis and reports on air quality, while also providing tools to visualise historical and real-time data.

3. DATA AND PRE-PROCESSING

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).
- Nitrogen dioxide (NO2) is a pollutant emitted mainly by vehicle traffic;
- Ozone (O3) is the main smog pollutant.

The dataset comprises four files, one for each pollutant in which the following measurements are reported:

- station_eu_code: European station code
- id_regione: numerical identifier of the region
- id_provincia: numerical identifier of the province
- id_comune: numerical identifier of the municipality
- station_code: national station code
- Regione: Region name
- Provincia: Province name
- Comune: Municipality name
- nome_stazione: Station name
- tipo_zona: Type of zone as classified according to Legislative Decree 155/2010
- tipo_stazione: Type of station as classified according to Legislative Decree 155/2010
- TIPO: Union of zone type and station type
- Lon: Longitude
- Lat: Latitude
- yy: year
- n: number of valid data available
- sup25: days on which the 25 µg/m^3 threshold was exceeded
- sup15: days on which the 15 µg/m^3 threshold was exceeded
- media_yy: annual average µg/m^3
- minimo: minimum value µg/m^3
- massimo: maximum value µg/m^3

To establish air quality on the basis of previous pollutants, reference is made to the European Environment Agency (EEA), which provides the following guidelines (Figure 1):

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

Figura 1. European Air Quality Index

3.1. Pre-processing

When pre-processing the data, I selected only the most relevant information, such as: station_eu_code,Regione, Provincia, Comune, Lon, Lat, yy, media_yy, minimo e massimo. Then I divided the data of the four agents by years, creating 12 different files, one per year, from 2010 to 2022. Thus obtaining 48 different files divided into 4 folders. I also grouped the different agents into one file, showing the annual averages for each agent, again divided by years.

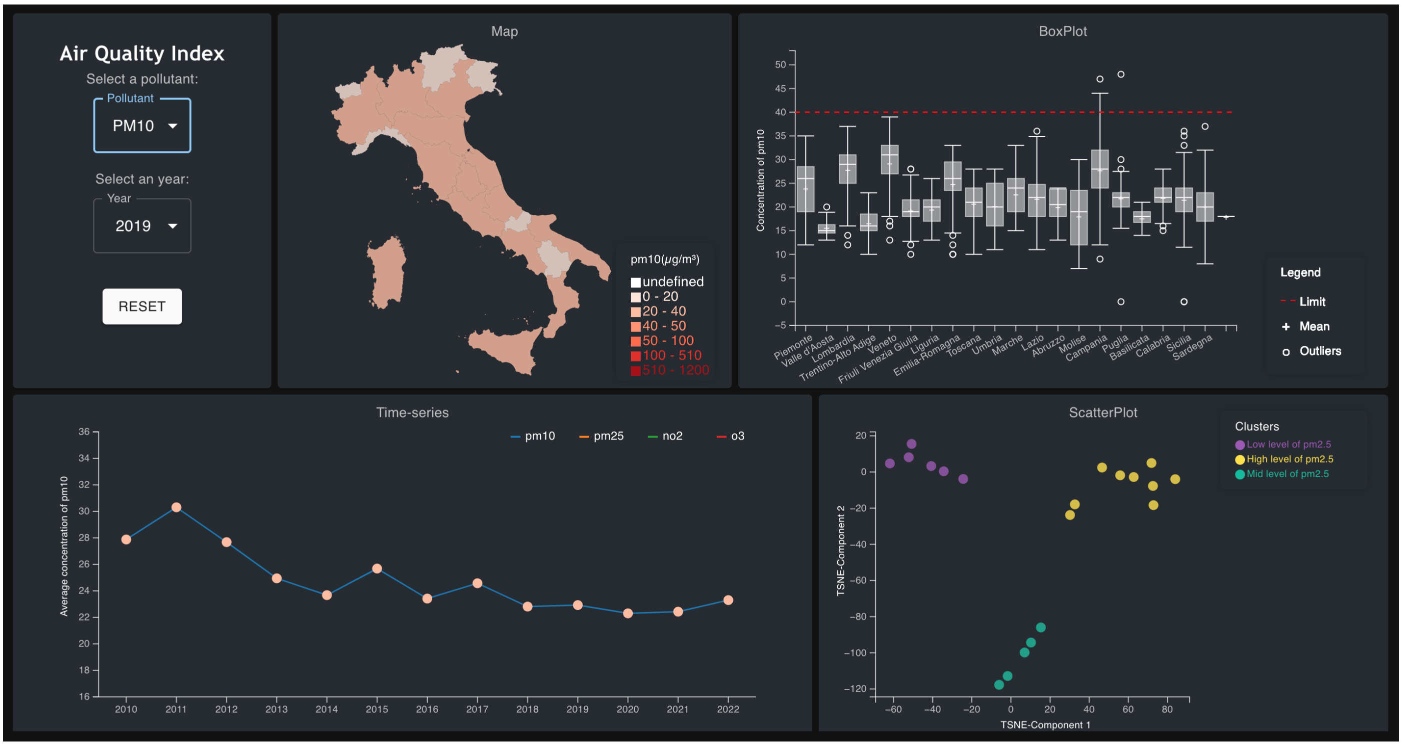

Figura 2. Tool

4. VISUALIZATION AND INTERACTIONS

The Visual Analytics tool for Air Quality Index [Figure 2] is constituted by four visualization, with which the user can interact.
The system includes:

- Map: Map of Italy divided into regions, depicted with a scale of color according to the mean concentration of different pollutant.
- Box-plot: it represents the array of all the measurements of a region, reported the median, the average, the maximum value, the minimum value and the outliers
- Time-Series: It analyze mean concentration trends over time.
- Scatter-plot: it report the results of t-SNE analysis reporting cluster of region with specific features.

There are three global interaction. The first one is a selection menu [Figure 1] (on the top left) with which the user can select a specific pollutant. The options are: PM10, PM2.5, NO2, O3 and Total. By selecting the 'Total' option the user can view the general situation of air quality due to the four pollutants. While for other options the user displays the different concentrations of individual agents. The second one is another selection menu with which the user can select a specific year. The years range from 2010 to 2022. If a user clicks on a region or on a box, they will be shown referring to that region.By clicking the 'reset' button the user can return the display of data to the global situation. Some views are interlinked, e.g. if you click on a region of the map or box-plot, the data for that region are also displayed in the box-plot/map and time-series. If a year is selected in the time-series, the data are also updated in the map, box-plot and scatter plot and in the general menu. The detailed explanation of single visualization and interaction is described below

4.1. Geographical Map

The Map [Figure 3] allow to show the mean concentration of the different pollutants in the Italy's regions. It consist of a map divided into 20 regions, each colour according to a shade of a red scale, according to the concentration of the selected pollutant. The more intense the color, the more concentration is greater, according to the legend. When the 'Total' option is selected, the color scale indicates the quality category of the air. The categories are Good, Fair, Moderate, Poor, Very Poor, Extremely Poor. This assignment is given based on the most polluting agent. So if for example Agent PM10 has a concentration classified as Good, Agent PM2.5 as Fair, Agent NO2

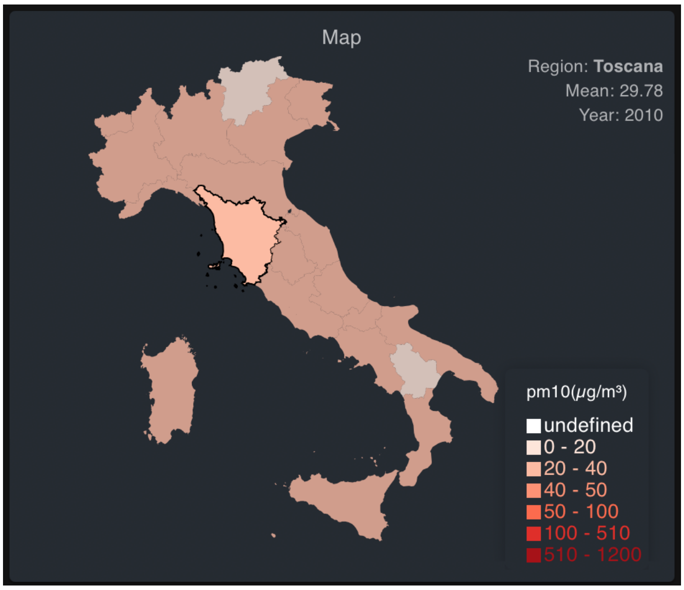

Figura 3. Map

as Fair, and Agent O3 as Good, generally the air quality of that region will be 'Fair', the worst one.

4.1.1. *Interactions of Map*

Mouse over region

When the mouse is placed over a region this is emphasised and a side tooltip appears (top right of [Figure 3]) showing the region name, the average of the selected pollutant and the selected year. If the selected option is 'Total' it shown the average of each agent for that region, the most polluting agent and the selected year is shown.

Click on region

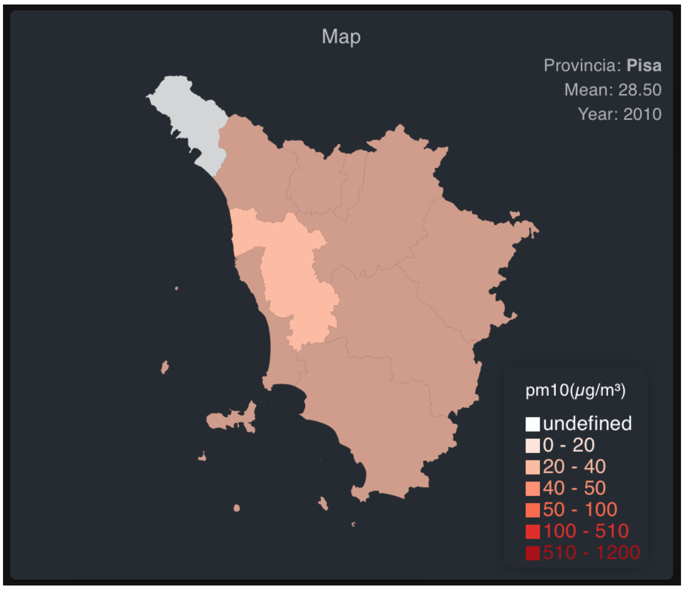

Figura 4. Click on region

When a region is clicked, the map is zoomed in and shown the region in detail. [Figure 4] In particular, the map is shown divided by provinces, also colored according to the average concentration of a selected pollutant or according to the general categorization.

Mouse over province

When the mouse is placed over a province this is emphasised and a side tooltip appears (top right of [Figure 4]) showing the province name, the average of the selected pollutant and the selected year. If the selected option is 'Total' it shown the average of each agent for that province, the most polluting agent and the selected year is shown.

4.2. Box-Plot

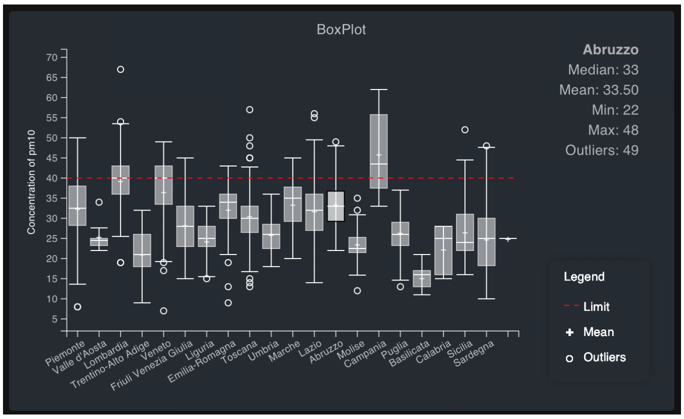

Figura 5. Box-Plot

The box-plot [Figure 5] depicts the data for each region and allows comparison of data from all regions, showing the mean, median, maximum and minimum values and outliers. The graph also shows a threshold indicating the European limit allowed for that pollutant. For the 'Total' option, the box-plot is not available (if selected, it represents the last agent shown).

4.2.1. *Interactions of Box-Plot*

Mouse over box

When the mouse is placed over a box this is emphasised and a side tooltip appears (top right of [Figure 5]) showing the region name, the median of the selected region,the mean, the max and min value and the outliers.

Click on box

When a box is clicked, the chart is zoomed in and shown the data of region in detail. [Figure 6] In particular, the zoomed box-plot show data divided by provinces.

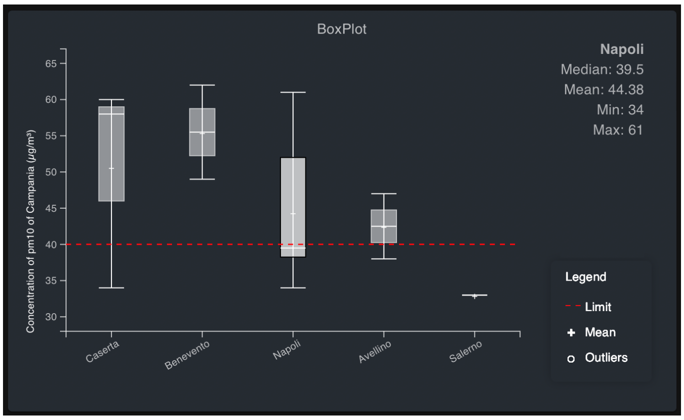

Figura 6. Click on box

Mouse over province
When the mouse is placed over a box representing a province this is emphasised and, at the same way of before, a side tooltip appears (top right of [Figure 6]) showing the province name, the median of the selected province,the mean, the max and min value and the outliers.

4.3. Time-Series

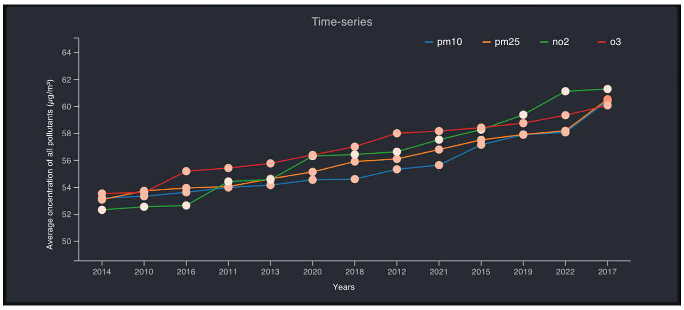

Figura 7. Time-series

The Time Series [Figure 7] depicts how the average concentration in Italy in general has changed over the years. On the x-axis we have the different years taken into consideration, from 2010 to 2022, on the y-axis the average of all regions. If the 'Total' option is selected, all four agents are shown together, divided into four different colours, as indicated by the legend.

4.3.1. *Interactions of Time-Series*

Selected region
When a region is selected, either in the map or in the box-plot, the time-series data are updated with those of the selected region, and the name of the region appears as the y-axis label.

Click on a dot
When a point representing a year is clicked on, it is selected and the data in the other graphs change according to the chosen year.

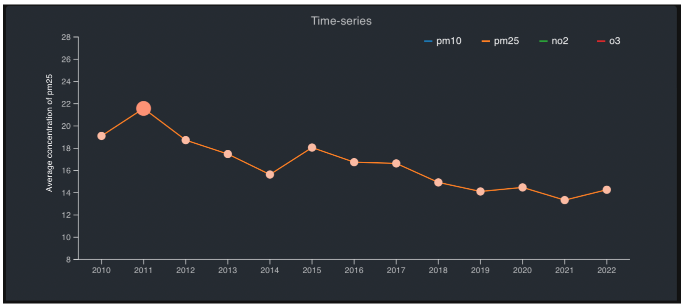

Figura 8. Time-series

4.4. Scatter-plot

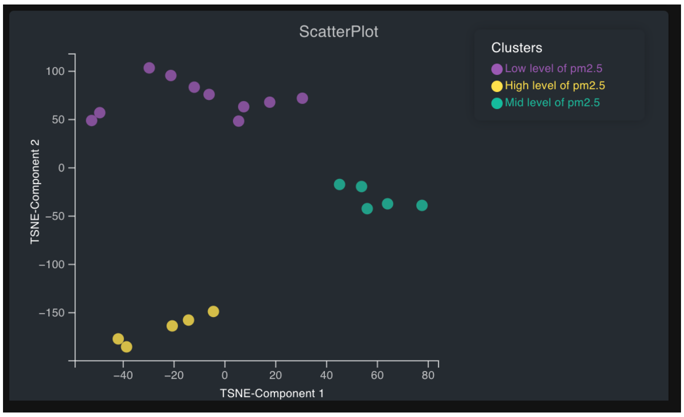

Figura 9. Scatter-plot

The Scatter-Plot [Figure 9] is the result of dimensionality reduction using t-SNE. This is a two-dimensional representation of all four agents where each point is associated with a region. The two components represent the differentiation of the points, so the more similar two regions have similar agent characteristics, the closer they will be to each other. It is therefore possible to identify three different clusters differentiated by three different colours. These colours are described in the legend that explains their meaning.

4.4.1. *Interactions of Scatter-plot*

Mouse over a dot
By hovering the mouse over a point, you can see the name of the region associated with it.

Change year
When a year other than the current year is selected, the data is updated with the new dataset.

5. ANALYTICS

The analytical process of the project focuses on applying t-distributed stochastic neighbor embedding (t-SNE) as a technique for dimensionality reduction. This process is developed using Python, with the support of the scikit-learn library for dimensionality reduction, clustering, and data preprocessing. The following steps were undertaken to achieve the objective:

- Data Preparation: The dataset from 2010 to 2022 was loaded using the pandas library. The dataset includes several air quality indicators such as NO2, PM10, PM2.5, and O3. A subset of these features was selected for further analysis. The mean values of these indicators were computed for each region and any rows with missing data were excluded.

- Standardization: The selected features were standardized using the 'StandardScaler' from the scikit-learn library to ensure that each feature contributes equally to the distance calculations during the t-SNE transformation.

- Dimensionality Reduction with t-SNE: The t-SNE algorithm was applied to the standardized data to reduce the dimensionality from four to two components. The t-SNE parameters were set with a perplexity of 2 and a random state of 42 to ensure reproducibility of the results.

- Clustering with K-Means: To group the regions based on the t-SNE results, the K-Means clustering algorithm was employed. The number of clusters was set to 3, and the algorithm was initialized with a random state of 42 for consistency. The clusters obtained from the K-Means algorithm were added to the dataset.

- Visualization: The results of the t-SNE transformation and clustering were visualized using a scatter plot. Each point represents a region, colored according to its cluster label. The regions were also annotated on the plot for better identification. The visualization was created using the seaborn and matplotlib libraries.

- Saving the Results: The transformed data, including the t-SNE components and cluster labels, were saved to a CSV file for report it also in the tool.

6. DISCOVERED INSIGHTS

In this section, the focus is on the insights derived from the analysis of the whole system. By analysing the data and results, it is possible to understand which areas are the most problematic. It is also possible to make a comparison between different years and understand which events may have influenced the results.

6.1. Decrease in pollutants

The first result that comes to light from the time-series analysis is that all agents (except O3) decreased in concentration from 2010 to 2022. This is because since 2010, Italy has implemented increasingly strict policies to reduce pollutant emissions in accordance with European directives, such as Legislative Decree 155/2010. These regulations have set precise limits for PM10, PM2.5 and NO2 emissions, imposing stricter controls. Local authorities have introduced Action Plans to improve air quality, including restrictions on vehicle traffic, reduction of industrial emissions and promotion of renewable energy. These actions have led to lower concentrations of pollutants. Tropospheric ozone (O3) behaves differently from other pollutants, as it is a secondary pollutant formed from chemical reactions in the atmosphere under the effect of solar radiation. Despite reductions in primary emissions, changing climatic conditions and rising temperatures have favoured the formation of ozone in the atmosphere, especially in warmer seasons. This explains why ozone concentrations have not decreased as much as those of other pollutants and, in some cases, may even have increased.

6.2. Most problematic regions for PM10 and PM2.5

The Italian regions of Campania, Lombardy and Veneto experienced high levels of PM10 and PM2.5 from 2010 to 2022 due to a combination of factors related to vehicle traffic, industrial activities, geographical conditions and domestic heating practices. Here is a more detailed analysis of the main reasons why these regions experienced high concentrations of fine particulate matter:

Vehicle traffic
Lombardy and Veneto: These two regions are part of the Po Valley, an area characterised by high urbanisation and intense road traffic. Large cities such as Milan (in Lombardy), Venice (in Veneto), and nearby cities such as Bergamo, Brescia, Verona and Padua have very high traffic volumes. Vehicle traffic, particularly from diesel engines, is a primary source of PM10 and PM2.5. Campania: The city of Naples and its metropolitan area suffer from traffic congestion and an older and more polluting vehicle fleet than the Italian average. Diesel vehicles and motorbikes contribute significantly to the emission of fine particulate matter.

Industry and production activities
Lombardy and Veneto: These regions are highly industrialised, with production activities that emit high quantities of air pollutants. Sectors such as chemicals, metallurgy, textiles and the production of building materials release particulate matter directly or through combustion processes. Campania: Industrial emissions are also significant in Campania, particularly in the Naples and Caserta areas, which host heavy industries and port activities that contribute to particulate pollution. In addition, the presence of construction sites, factories and power plants in the region aggravates the situation.

Unfavourable geographical and meteorological conditions
Po Valley (Lombardy and Veneto): The Po Valley is one of the most polluted areas in Europe due to its geographical conformation. Being surrounded by mountains (Alps and Apennines), the plain suffers from poor ventilation and meteorological phenomena such as thermal inversion that trap pollutants close to the ground, especially in winter. This causes an accumulation of particulate matter in the air, worsening air quality significantly.

Illegal burning of waste
Campania: A specific problem in this region is the illegal burning of waste, particularly in the so-called Terra dei Fuochi (located between the provinces of Naples and Caserta). Uncontrolled combustion of plastics, chemicals and municipal solid waste releases large quantities of particulate matter into the air, contributing to increasing PM10 and PM2.5 levels and further worsening air quality.

6.3. The contradictory situation of the covid-19 pandemic

In 2020, during the COVID-19 pandemic, despite the lockdown and the reduction of economic activities and traffic, higher concentrations of PM10 and PM2.5 were recorded in some areas of Italy. This phenomenon may seem contradictory, as a decrease in air pollution was expected, but there are several reasons that explain this situation. During 2020, weather conditions in some parts of Italy were not favourable to the dispersion of pollutants. Prolonged periods of high pressure, with little wind and thermal inversion, have trapped pollutants close to the ground. These phenomena, especially in the winter months, limit air exchange, leading to the accumulation of particulate matter in the atmosphere. During the lockdown, many people stayed at home for long periods, which led to an increase in domestic heating, particularly in rural and remote areas where fuels such as wood, pellets and other biomass-based materials are used. Biomass combustion is a significant source of particulate matter (both PM10 and PM2.5).

7. APPLICATION AND UTILITIES

This section highlights the practical impact of the Visual Analytics tool for the Air Quality Index, targeting its intended users and illustrating specific use cases to demonstrate its utility.

7.1. Intended users

The Visual Analytics tool for the Air Quality Index is designed to be a valuable resource for environmental analysts, policymakers, and urban planners. By providing comprehensive and interactive visualizations of air quality data, the tool aids these users in understanding pollution patterns and making informed decisions to improve public health and environmental conditions.

1. Environmental Analysts: These professionals can use the tool to monitor and analyze trends in air pollutant levels across different regions. The tool's capability to visualize historical data and detect anomalies supports analysts in identifying areas with critical pollution levels

and investigating underlying causes.

2. Policymakers: The tool assists policymakers in formulating and evaluating environmental policies. By leveraging data insights, policymakers can develop targeted interventions to reduce emissions and enhance air quality. The interactive features of the tool allow for scenario analysis, enabling the assessment of potential outcomes of proposed measures.

3. Urban Planners: Urban planners can utilize the tool to design and implement strategies for sustainable urban development. The spatial distribution of pollutants, visualized on the map, helps planners to identify high-risk areas and prioritize actions such as green space development, traffic management, and industrial zoning to mitigate pollution impacts.

7.2. Use Cases

In this subsection, two distinct use cases are presented. These real-world scenarios demonstrate the versatility of the Air Quality Monitoring system and highlight its practical applications in enhancing environmental quality and public health.

1. *Urban Planning and Development*

   Scenario: As part of a new urban development project, a city planner decides to use the Air Quality Monitoring system to ensure that the new infrastructure will not adversely affect air quality. The planner accesses historical and real-time data to identify areas with high pollution levels and determine the sources of these pollutants.

   Process: The planner examines the geographical map to visualize pollution concentrations across different regions. By selecting specific pollutants, such as NO2 and PM10, the planner identifies areas with high levels of vehicular emissions. The time-series data is then analyzed to observe trends and seasonal variations in pollutant levels.

   Outcome: Utilizing these insights, the planner collaborates with transportation and environmental experts to design traffic flow patterns and green spaces that reduce pollution exposure. The system's predictive capabilities allow the planner to assess the potential impact of the new development on air quality and make necessary adjustments to the project plan, ensuring a healthier urban environment.

2. *Public Health Intervention*

   Scenario: During a severe air pollution episode, a public health official uses the Air Quality Monitoring system to develop an intervention plan aimed at reducing health risks for vulnerable populations. The official accesses real-time data to identify the most affected areas and the pollutants contributing to the poor air quality.

   Process: The official analyzes the box-plot to identify regions with the highest median pollution levels and uses the scatter-plot to visualize the relationship between different pollutants. The time-series data helps in understanding the duration and intensity of the pollution episode.

   Outcome: Based on this analysis, the official issues health advisories for residents in the affected areas, recommending measures such as staying indoors and using air purifiers. The official also collaborates with local authorities to implement emergency measures, such as restricting vehicle use and suspending industrial activities. The intervention plan, supported by the system's data-driven insights, helps mitigate the health impact of the pollution episode.

8. CONCLUSION & FUTURE IMPROVEMENTS

The Visual-Analytics Air Quality Tool systems represents a significant advancement in environmental monitoring and public health protection. By providing detailed data on air quality, the system empowers various stakeholders to make informed decisions and implement effective strategies to combat air pollution. The system's analytical capabilities, including geographical mapping, time-series analysis, and clustering, offer valuable insights into pollution trends and sources. Through practical applications such as urban planning and public health interventions, the system demonstrates its versatility and utility in enhancing environmental quality and safeguarding public health. The data-driven approach facilitates proactive measures, ensuring timely and targeted responses to air quality issues. In conclusion, the Air Quality Monitoring system is an indispensable tool in the fight against air pollution, contributing to the creation of healthier and more sustainable environments. By leveraging the system's capabilities, stakeholders can work collaboratively to address the complex challenges of air pollution and achieve long-term improvements in air quality.

Continuing technological advances represent an important opportunity for ongoing research and developments. In particular, emerging technologies, such as IoT (Internet of Things) devices and advanced sensors, could contribute to a more comprehensive and real-time data collection process. This upgrade not only promises to improve the investigation capabilities of the methodology, but also paves the way for more dynamic and responsive trace management strategies. Furthermore, future efforts could focus on refining the design to handle additional variables or integrate advanced techniques for more accurate predictions.

REFERENCES

BACHECHI, C. et al. Visual analytics for spatio-temporal air quality data. In: IEEE. *2020 24th International Conference Information Visualisation (IV)*. [S.l.], 2020. p. 460–466.

EEA. Disponível em: <https://www.eea.europa.eu/en/topics/in-depth/air-pollution>.

LIAO, Z. et al. A web-based visual analytics system for air quality monitoring data. In: *2014 22nd International Conference on Geoinformatics*. [S.l.: s.n.], 2014. p. 1–6.