EnVis: visualization tool for environmental quality

Laura Papi
papi.1760732@studenti.uniroma1.it

Francesca Restante
restante.1745106@studenti.uniroma1.it

Abstract - **Environmental pollution poses significant challenges to public health and the sustainability of ecosystems. In order to address these concerns, there is a growing need for comprehensive monitoring and analysis of air and water quality, particularly in urban areas where pollution sources are concentrated. This paper presents EnVis, an Environment Visualization System specifically developed for studying air and water quality in multiple cities across Italy. By integrating diverse data sources, including pollutant measurements, waste management records, industrial activities, and noise pollution levels, EnVis provides a better understanding of the environmental factors influencing urban quality.**

Index terms - **Air quality, Water quality, Visual Analytics, Environment**

INTRODUCTION

The degradation of air and water quality has become a pressing environmental issue in many regions worldwide, and Italy is no exception. Urban centers, in particular, face complex challenges due to industrial activities, waste generation, and transportation networks.

The goal of this project is to build a data visualization tool for the quality of the environment in Italy.

Environmental quality is obviously a general term, that can include a wide variety of factors, the ones that we take into consideration for the purposes of this project are:

- Air and water pollution
- Presence of green areas
- Types of circulating vehicles (divided by fuel category)
- Accessibility to public transport
- Industrial and commercial noise
- Waste management

The visualization system that we propose to implement has the purpose of providing a graphical representation of all these factors, allowing the user to immediately grasp the quality of the environment in which he lives, simply by observing the graphs and eventually interacting with them through simple and intuitive operations.

Also, the plots will provide several ways to compare cities, letting the user find out which are the cities that rank best in the factors that he values the most.

At last, by implementing many visualizations that provide comparisons between the different factors, it is possible to gain insights on how the various factors influence each other, and therefore to discover how very heterogeneous variables contribute to defining the quality of the environment we live in.

1. VISUALIZATION AND INTERACTION

The first design choice was to allow the user to select the target city.

The system is then articulated in 5 different graphs, with which the user can interact.

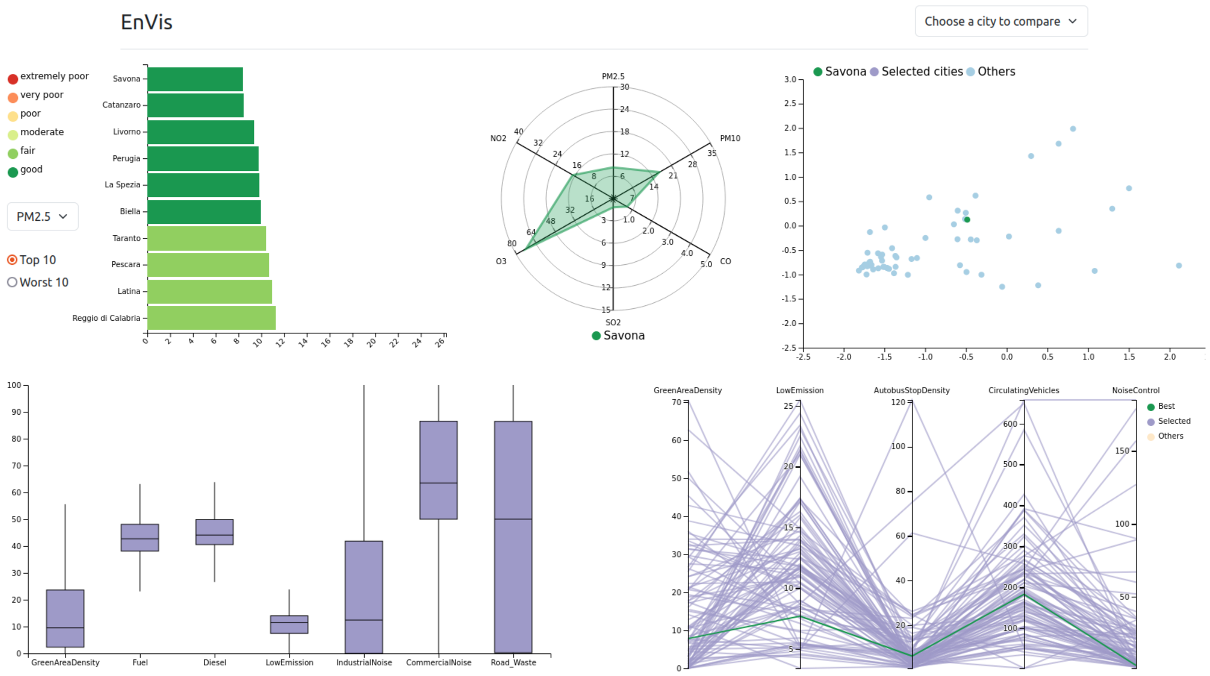

Figure 2

The first step towards the realization of the system was to design a mockup of the overall visualization [Figure 1], which then was implemented to obtain the final working system [Figure 2].

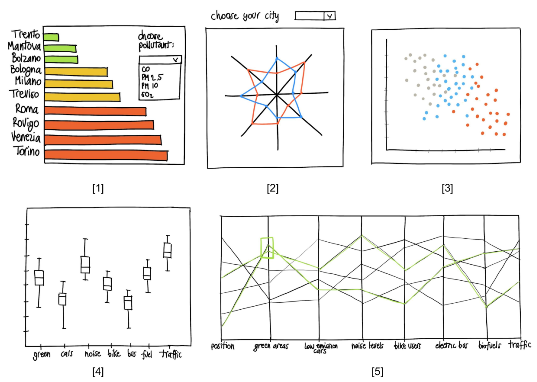

Figure 1

The design idea of these 5 visualizations is described in detail through the following chapter.

1.1 Air Quality Ranking

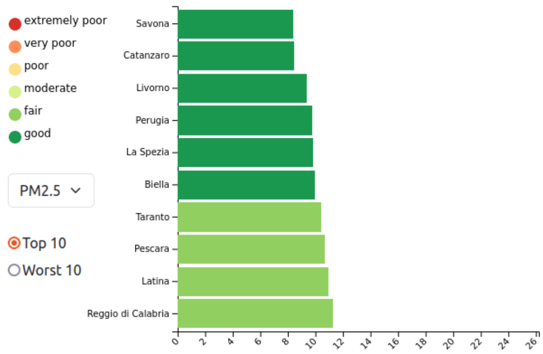

Figure 3

The first view [Figure 3] focuses on air quality, it provides a bar chart that ranks cities based on the level of pollution of the air.

To determine what the air quality is, we chose to stick to the European Environment Agency definition of Air Quality Index (AQI)[11], which takes into consideration five air pollutants and classifies the air using the corresponding legend [Figure 4].

For the colors, we have chosen not to use those of the original legend, but to use a 6 colors divergent scale from the ColorBrewer website[9].

In this visualization the user can choose which pollutant to analyze, and the cities will be ranked with respect to that pollutant, the color scale for this plot is built to correspond to the official AQI classification [Figure 4][11].

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
Sulphur dioxide (SO_2)	0-100	100-200	200-350	350-500	500-750	750-1250

Figure 4

To these five pollutants that take part of the European AQI, we also chose to display the CO levels, which is part of the U.S. Air Quality Index definition[13].

1.2 Air Quality Benchmark

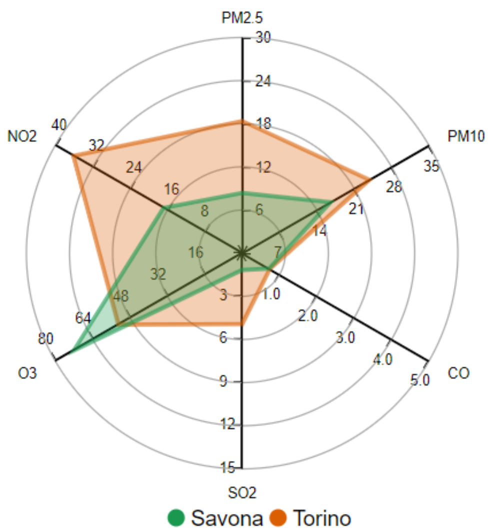

Figure 5

The second visualization [Figure 5] is strictly linked to the first one. The source data is still the pollutant values per each city, but instead of just ranking the cities, thanks to the radar chart representation, it allows the user to get an overview of the currently selected city over all the 5 pollutants included in the AQI.

Also, it allows to display in the same plot both the currently selected city (the gray one), and the best ranked city in the current selection (the green one).

The radar chart is a great visualization tool for this kind of benchmark, because it is able to both summarize the data of a single city (by displaying the area extent and all the values for all the parameters), and also provide an immediate way to compare it with another city. So the user can get an idea of which city is altogether the best one, but also in which parameters they differ.

1.3 Water Quality Analysis

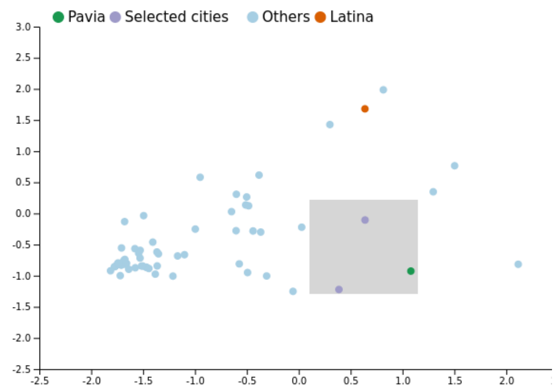

Figure 6

With the third chart [Figure 6] the water quality factor is introduced. In particular, we define water quality based on pesticide contamination levels of the surface waters. Since water monitoring is done by checking many different substances, and in order to be able to interpret them, a high level chemical knowledge is needed, we opted for a scatter plot representation, that doesn't directly reports the specific value for each monitored pesticide, but creates a bidimensional visualization, that through the PCA algorithm, is able to display all the cities based on the two components that best differentiate them.

With such a representation it is possible to see which cities are similar, and which are outliers.
We can even identify clusters of cities that share similar values for one or both the pca variables.

This view also allows interaction, the user can brush areas of the graph, selecting groups of cities, or he can simply move the mouse over a dot to find out which city it is.

When the user brushes a section of the scatterplot, all the other plots will be updated showing the results of the computations over the new set of cities chosen by the user.

This graph also highlights the best city of the current selection, and the city chosen by the user (in green and red respectively, using divergent color scales from the ColorBrewer website[9]).

1.4 Secondary Factors Statistics

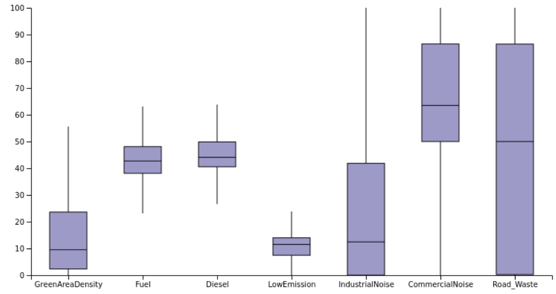

Figure 7

In the fourth visualization we have a box plot [Figure 7] in which we take into consideration different factors concerning the lifestyle of the cities . In particular, we examine green area density, petrol and diesel distribution, low emission vehicles, industrial and commercial noise and the road waste collection. On the y axis we measure the percentage of each factor. Through this plot we have a summary of the statistical values, especially the median useful for considering the trend of a set of cities.

1.5 Secondary Factors Analysis

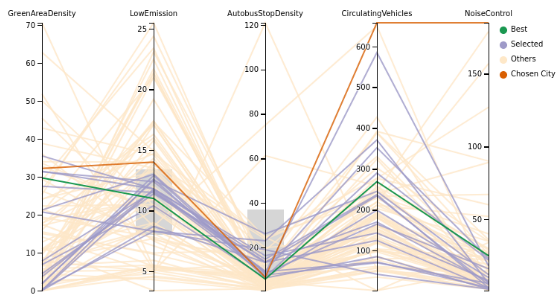

Figure 8

In the last one we have a parallel coordinate plot [Figure 8] useful for visualizing multidimensional data through different axes.
Each axis could represent a factor we want to analyze in detail (related to the one precedently analyzed in the box plot), while each line corresponds to a city that

intersects each axis at its corresponding value.
To identify a single city there is a tooltip function called when the mouse is over the line that provides the corresponding name.
Due to a concentration of lines in particular areas of the plot it is difficult to distinguish between the cities.
The city that highlighted with the red color is the one that in the current selection has the best parameters for air quality
There is a brushing feature on all the axes that helps to filter and highlight only certain cities.

1. DATASET AND PREPROCESSING

As we have seen in the previous chapters, there are many factors that we have chosen to include in this visualization system, and this inevitably leads to having many different datasets, from many different sources.

We briefly describe which sources we used for each view, and how we had to integrate the different datasets.

2.1 Air Quality Data

This data was needed to implement both the first and the second visualizations.
As mentioned above, the source was the European Environment Agency[1], which provides an online tool for filtering the data and downloading only the chosen rows and columns.
In particular we downloaded only Italy's data from 2021.

Since this dataset provides data for each monitoring station in Italy, we had to preprocess the data (using Pandas library[5]) to compute the mean value for each city and for each pollutant.
At the end of preprocessing we obtain a table with the form [Air Pollutant, City, Air Pollution Level].

2.2 Water Quality Data

As described above, water data is needed to implement the third view.
The dataset regarding water pollution was retrieved from the "ISPRA - Portale Pesticidi" download section[2].

Similarly to the AQ data, the monitoring is documented by each sampling station.
This again means that we had to compute the mean during a preprocessing step.
But also, in this case, the data provided by ISPRA is divided by region, so we also had to merge the dataset from each region into a national one.

2.3 Secondary Factors Data

The data used for this visualization were extracted respectively from the ISTAT website[3], from the tables related to each factor.
Secondary factors data play an important role in providing a comprehensive understanding of the overall environmental conditions.
Here's how these secondary factors influence a visual analytics system:
- <u>Presence of green areas</u>:
It can help identify areas with better air quality due to the presence of green spaces and highlight the need for additional green areas in polluted zones.
- <u>Types of circulating vehicles</u>:
It can analyze areas with high levels of vehicle emissions, highlight traffic

hotspots, and support policy decisions to promote cleaner and more sustainable transportation options.

- Accessibility to public transport:
A well-developed public transport system can help reduce private vehicle usage, leading to lower emissions and improved air quality.
- Industrial and commercial noise:
Noise pollution can disrupt aquatic ecosystems and lead to increased stress levels in humans.
- Waste management:
Inadequate waste disposal can lead to air pollution from incineration or open burning and water pollution from improper waste disposal sites.

2. ANALYTICS

For the analytics the main components are represented by the boxplot, which do a dynamical computation of statistical factors such as:

- Median: the middle value of the dataset when it is sorted in ascending order. In our boxplot, the median is represented by a line. It provides a measure of the central tendency of the data.
- Quartiles: divide the dataset into four equal parts. The first quartile (Q1) represents the lower 25% of the data, while the third quartile (Q3) represents the upper 25% of the data. The interquartile range (IQR) is the difference between Q3 and Q1
- Whiskers: The whiskers of a boxplot represent the range of the data beyond the quartiles.

Upper Whisker: Q3 + (1.5 * IQR)
Lower Whisker: Q1 - (1.5 * IQR)

3. LITERATURE REVIEW

In recent years, awareness of pollution has definitely grown, and many studies have been conducted in order to predict and control the future trend, but also to better analyze and understand the current state. One of the primary concerns is air pollution, which can be defined as “the presence of substances in atmosphere in sufficient concentrations so that they threaten to be harmful to human, plant, animal life, and the most drastic damage to the environment and the climatic conditions.”[10]. So, for a substance, in order to be considered as an air pollutant, there should be proven hazardous effects on human health or environment, such as for the substances listed and described in the cited study[8] [Figure 9].

POLLUTANTS	EFFECT ON HUMAN HEALTH	EFFECT ON ENVIRONMENT
Sulphur Dioxide (SO_2)	• Respiratory problems. • Causes irritation to the eyes and skin. • Lung cancer. • Cardiovascular diseases and mortality rate.	• Damage vegetation. • Causes Corrosion. • Acid Rain. • Plant and water damage. • Aesthetic damage.
Nitrogen Dioxide (NO_2)	• Infections in respiratory system. • Capacity of carrying oxygen in the blood decreases. • Problems in thyroid gland. • Asthma.	• Aquatic ecosystem affected. • Ozone layer depletion
Carbon Monoxide (CO)	• Heart diseases and chest pain. • Causes flu.	• Concentration of greenhouse gases increases. • Burning of biomass.
Ozone (O_3)	• It causes problems related to the respiratory system. • Causes lung problems. • It also causes irritation in throats.	• Crop yield decreases. • Growth of forest decreases.
Particulate Matter (PM)	• Cancer and respiratory illness. • Bronchitis, asthma, pneumonia etc.	• Deposition of acid in air and water. • In nutrient cycles, changes occur. • Ecosystem does

Figure 9

These are in fact the pollutants that we take into consideration for our AQ visualizations.
It's important to note that in these studies Particulate Matters are all considered as a single type of pollutant, instead we have chosen to treat PM2.5 (diameter << 2.5μm) and PM10 (diameter << 10μm) as two distinct pollutant factors (distinction that we also find in the EEA's AQI definition[11]), because PM2.5 is considered to be fairly more dangerous, and its limit values are stricter[12].

Regarding the visualization of the pollution data that we described, several studies can be found that assess the necessity for a way to visualize these large scale datasets.

In particular AirVis[6] finds a possible solution for this problem, but it mainly focuses on how the air pollutants propagate, and the intended users for the presented system are domain experts, that thanks to these visualization systems would be able to interpret and infer pollution propagation factors.

Instead, what we are interested in providing, is a way for a non-expert user to grasp information and gain insights on pollution in its geographical area, without the need for a deep comprehension of the scientific phenomena.

A study similar to what we aim to obtain, is the visualization system proposed in [14]. It has the purpose to make complex data (about air pollution) understandable by exploiting the intuitiveness of graphical representations.

But they chose to focus on spatio-temporal data analysis, providing a very in-depth understanding of the air pollution phenomena.
While we were more interested in providing more variegated information, not only focusing on air pollution, but also including and analyzing the relations with other environmental factors such as water pollution, traffic, and all the others described in the previous chapters.
So, we chose to forgo the time dimension, instead focusing on representing the relationships between all the different factors.

4. CONCLUSIONS

Our visualization system is designed to provide to the intended audience a means for analyzing their city and other cities to make a comparison in terms of environmental context.

General insights

The Air Quality chart [Figure 3] and the Statistic chart [Figure 7] offer an overview of all the cities in Italy.
The bar chart allows the user to immediately identify the city with the best air quality (the default pollutant considered is PM2.5, which is the most hazardous).
Also, if we choose to display the 10 worst cities instead of the best, we can notice that even the worst city still has acceptable values of the PM2.5 concentration. So, we can assume that the overall quality of the air in Italy is good enough to not cause any health issues.

From the box plot we can notice that the percentage of the diesel vehicles have a very high mean, even higher than the petrol one. Also the values of the commercial noise are much higher than the industrial ones, suggesting that most cities aren't affected by the noise produced by industries.

Intended user

The target user of this visualization system is a non-expert that is mainly concerned about the quality of life of the city he lives in or wants to live in. For instance, if I found out that the PM2.5 levels in my city are above the limits, I must act accordingly to prevent health issues, for example using a mask with filters.

Use cases

We provide two examples of use for which our system can be useful.

1. *The user wants to find which is the best city where he can live according to his needs and priorities [Figure 10]*

First of all the user might want to select the city in which he currently lives in, in order to compare it to the other cities. Suppose the user selects Roma for the comparison.

The radar chart provides a very intuitive benchmark from which we can immediately find out that the PM2.5 and PM10 levels are considerably higher in Rome, which could be a concern for our user.

The scatter plot, which represents the similarity in terms of water pollution, shows Roma and Savona as almost identical, so if the user is satisfied with the water quality in Roma, he might consider moving to Savona, since he would have a better air quality, without giving up to quality water.

Lastly, the parallel coordinate plot is the most important for this use case, as it allows the user to filter the cities based on his personal needs. For example he might want to live in a city with many parks (selects only lines above Roma in the GreenArea column), then he selects only the cities with a few circulating vehicles, and a high percentage of low emission vehicles.

The cities that match all these requisites are proposed by the system as cities in which the user might want to move to. In particular in this example the user finds 4 compatible cities, among which Biella is the best ranked one for its air quality. Therefore, our system suggests to the user that Biella would be the ideal city to meet all his needs, according to the available data.

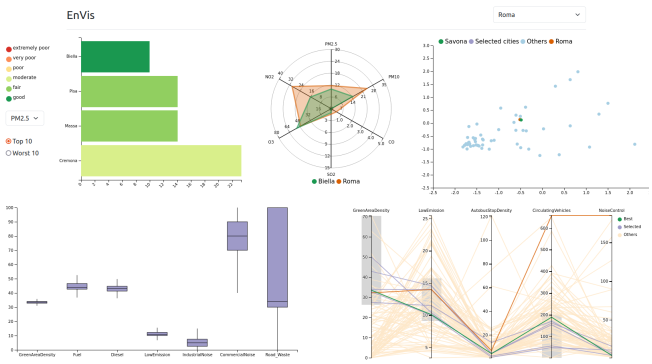

Figure 10

2. *The user wants to find out what are the critical points of his city and consequently improve them*

Suppose that our user is not interested in moving, and instead he wants to

understand the main issues of the city he currently lives in, in order to improve it. For instance, we're assuming to choose Latina as the selected city.
It is immediate to notice that this city is in the top 10 in the bar chart, so this means that the quality of the air is good enough. From a detailed vision of the pollutant (emerging from the radar chart) we discover that it is actually very similar to the best city but the Particulate has a higher value. In fact from the parallel plot we can see that the levels of circulating vehicles and the autobus stop density are more elevated, so more pollution is created.
If we analyze the city based on the water quality, it is evident that there are no similarities with cities like Rome (which should have a good level of water quality) and Savona (best city). In the same cluster we can see that cities similar to Latina have more diesel vehicles (which are more polluting) than petrol ones.
On the other hand, the box plot shows that there is little green area compared to the general average.
So if the goal is to improve this city, the focus is on increasing green areas and replacing polluting vehicles with low-emission ones. In addition, the water quality is not part of the cluster that contains the best city, so surely there will be something to improve in terms of pesticides.

5. REFERENCES

[1] European Environment Agency, Annual AQ statistics. https://discomap.eea.europa.eu

[2] ISPRA, Portale Pesticidi. https://sinacloud.isprambiente.it

[3] ISTAT, Ambiente urbano https://www.istat.it/it/archivio/281184

[4] Michael Bostock, D3.js (Version 4.13.0) [Source Code]. https://d3js.org/

[5] Wes McKinney, Pandas (Version 0.25.3) [Source Code]. https://pandas.pydata.org/

[6] Z. Deng *et al.*, "AirVis: Visual Analytics of Air Pollution Propagation," in *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 1, pp. 800-810, Jan. 2020, doi: 10.1109/TVCG.2019.2934670.

[7] Y. -R. Zeng, Y. S. Chang and Y. H. Fang, "Data Visualization for Air Quality Analysis on Bigdata Platform," *2019 International Conference on System Science and Engineering (ICSSE)*, Dong Hoi, Vietnam, 2019, pp. 313-317, doi: 10.1109/ICSSE.2019.8823437.

[8] R. Gupta, N. Singh and B. Singh, "Neural network based GIS application for air quality visualization," *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*, Dehradun, India, 2016, pp. 169-172, doi: 10.1109/NGCT.2016.7877409.

[9] Cynthia Brewer, Mark Harrower and The Pennsylvania State University, ColorBrewer (Version 2.0) [Source Code]. https://colorbrewer2.org

[10] Niharika, Venkatadri M, and Padma S.Rao, A survey on Air Quality forecasting Techniques, vol. 5(1). International Journal of Computer Science and Information Technologies, 2014, pp.103-107

[11] European Environment Agency, Air Quality Index. https://www.eea.europa.eu

[12] Lu F, Xu D, Cheng Y, Dong S, Guo C, Jiang X, Zheng X. Systematic review and meta-analysis of the adverse health effects of ambient PM2.5 and PM10 pollution in the Chinese population. Environ Res. 2015 Jan;136:196-204. doi: 10.1016/j.envres.2014.06.029. Epub 2014 Nov 25. PMID: 25460637.

[13] AirNow, U.S. Air Quality Index. https://www.airnow.gov

[14] Li H, Fan H, Mao F. A Visualization Approach to Air Pollution Data Exploration—A Case Study of Air Quality Index ($PM_{2.5}$) in Beijing, China. *Atmosphere*. 2016; 7(3):35. https://doi.org/10.3390/atmos7030035