PolluViz: A Vizualization tool to view pollution data

Varshit Jain, Sourav Mangla, Deep Ruparel

Abstract—
PolluViz aims to explore the effective representation of multidimensional environmental and transportation data using various data visualization techniques. The project focuses on the importance of atmospheric variables, such as CO, NO2, O3, PM2.5, and PM10, and worker commute data in understanding environmental impacts and transportation patterns. With the integration of diverse datasets, the project highlights the significance of employing effective means of data visualization to convey multiple dimensions of data accurately. The project emphasizes the challenges of interpreting multidimensional data and the need to handle missing values in a transparent manner that does not mislead the viewer. Overall, this project contributes to the advancement of data visualization techniques to improve our understanding of the environment and transportation systems.

1 Introduction

Air pollution in the United States has many sources, ranging from transportation to industrial activity to power generation. It poses a significant threat to public health, and it is responsible for numerous respiratory and cardiovascular diseases. Unfortunately, despite the availability of data on pollution in various databases, there is no consolidated platform that provides comprehensive information on pollution in the United States. Along with that, the existing datasets also have several missing information like pollution in counties with small populations. PolluViz also aims to provide a solution to the problem of missing information. There is a well-established relationship between income and pollution, with several studies suggesting that higher levels of income are associated with lower levels of pollution. The exact nature of this relationship can vary depending on the specific context and types of pollutants being examined, but some general trends have been observed. One explanation for the income-pollution relationship is the environmental Kuznets curve (EKC) hypothesis, which suggests that pollution initially increases as a country develops and industrializes. [5]
Polluviz addresses this gap by providing an interactive platform that presents data on air pollution in the United States. By focusing on three key indicators of pollution, Polluviz offered users the ability to explore the relationship between pollution and population, income, and commuting. Geospatial mapping allowed users to view pollution data across different counties in the United States, providing a comprehensive picture of the pollution situation.
Polluviz provides important insights into the relationship between pollution and key indicators of human activity. For example, Polluviz can show how commuting patterns contribute to air pollution in certain areas or how pollution disproportionately affects low-income communities.
The interactive capability of our dashboard allows users to see how different variables interact with each other, providing a more nuanced understanding of the factors that contribute to pollution. Overall, Polluviz aims to fill an important gap in existing data sources related to pollution in the United States. By providing an interactive and intuitive platform that presents data on pollution across different counties, Polluviz offers users a comprehensive understanding of the pollution situation in the United States.

The aims of this research are:

- To develop an interactive visualization solution, called Polluviz, that collects and presents data on pollution in the United States. It focuses on three key indicators of air pollution: population, income, and commuting.
- To tackle the issue of missing values in a sophisticated manner, using a methodology that leverages the values of surrounding attributes to fill in gaps. By adopting this approach, the solution aims to optimize the accuracy and completeness of the data.
- Offers users a responsive interface to explore the relationship between pollution and key indicators mentioned above. It allows users to interact with the data and gain a comprehensive understanding of the pollution situation across different counties in the United States.

- *Varshit Jain: varshitjain@arizona.edu,*
- *Sourav Mangle : souravmangla@arizona.edu,*
- *Deep Ruparel : deepruparel@arizona.edu.*

2 Background

Atmospheric variables are fundamental physical and chemical properties of the Earth's atmosphere that exhibit spatiotemporal variability. Carbon monoxide (CO), nitrogen dioxide (NO2), ozone (O3), and particulate matter (PM2.5 and PM10) are among the most common atmospheric variables, where PM2.5 and PM10 are particles with diameters of 2.5m or less and 10m or less, respectively. Typically measured using the air quality index (AQI), these parameters can significantly impact human health and the environment, and are hence, subject to close monitoring by governmental agencies and research institutions.

Worker commute data, on the other hand, constitute valuable information that elucidates how individuals traverse from and to their respective workplaces. This data serves as a vital tool in understanding commuting patterns within a given region, recognizing transportation-related predicaments, and informing transportation planning and policy decisions. The dataset encompasses details regarding the number of workers who commute via distinct transportation modes, such as transit, carpooling, and walking, in addition to the distance and duration of their commutes. Furthermore, it provides insights into contemporary trends, such as telecommuting, as well as the influence of commuting on traffic congestion, air quality, and public health.

With the amalgamation of diverse datasets from varied sources, it becomes imperative to develop effective means of data representation. Multidimensional datasets pose a challenge in terms of interpretation as they entail more than two variables. Hence, data visualizations must employ techniques such as color coding, size, shape, and positioning to convey multiple dimensions effectively. Some effective methods for depicting multidimensional datasets include the bubble chart, scatter plot matrix, and parallel coordinates.

As regards environmental data visualization, line charts, bar charts, and scatter plots are typically utilized. However, these visualizations may not always be sufficient to communicate complex environmental data. Thus, interactive visualizations may serve as effective alternatives.

In cases of missing data, it is crucial to handle the missing values in a manner that does not mislead the viewer. One approach is to employ data imputation techniques to estimate the missing values, while another is to exclude them and display only the available data. However, this can lead to biased interpretations. A more transparent

approach involves including a disclaimer that acknowledges the missing data and its impact on the analysis.

2.1 Related Work

Conventional plots, such as scatter plots, are used to analyze time-series data and show the correlation among various factors in air pollution exploration [7] [18]. The rapid development of web technology allows interactive visualizations to combine various technologies [3], such as HTML for page content, CSS for aesthetics, JavaScript for interaction, and SVG for vector graphics. These technologies render sharing intuitive information highly convenient. Many spatial distribution explorations and complex representations of pollution [14] [16] rely on standalone or proprietary software products like ArcGIS. This software allows users to visualize multivariate statistical analysis intuitively, as well as explore and understand spatiotemporal and multivariate patterns. Nevertheless, this type of exploration does not consider the extendable and sharable functions that are important for the public to acquire multi-perspectives and latest information on air pollution. Numerous researchers focus on systematic theories of visual exploration for spatio-temporal data.

Kraak [10] claims that graphics can reveal patterns that are not necessarily visible when conventional map display methods are applied, demonstrating the usefulness of geovisualization. [5] provides a visualization of the air pollution in the different parts of Beijing, China using heatmaps and calendar view, and a few other static visualizations. Interactive dashboard is a kind of visual analytical method that concisely combines texts, images, charts, maps, videos, and gauges on a single screen to allow users'instant perception. The interactions on dashboards, such as selecting, filtering, searching, arranging, or drilling down, would additionally empower users with the flexibility to view and explore information effectively [8].

Map-based dashboards have the advantages of presenting the overview information, communicating the spatial complexity, and providing useful insights to various stake-holders . [12] designed a visual analytical tool with small multiples of maps and bar charts to show changes in events by linking these visualizations together.

The authors in this presentation [9] made a dashboard to evaluate the SDoH of the different neighborhoods in the state of Ohio and to show the SDoH they made use of eight health dimensions. This dashboard solution allows visualizing OCOI. This paper [17] shows the air pollution in the city of Lisbon and shows only one attribute of pollution which is PM25. The platform collects data using low-cost portable sensors and integrates it with other environmental contextual data to create an interactive dashboard.

[15] The researchers used GIS and LIDAR data to create detailed maps of pollution levels in different areas of Prague, which allowed for effective data visualization. However, the cost of data collection and modeling systems may make them prohibitive for local authorities. Additionally, while GIS and LIDAR data can provide detailed information about pollution levels in different areas, they may not capture all sources of pollution or account for all factors that contribute to environmental degradation. Therefore, it is important to use these tools in conjunction with other methods and to interpret the results with caution.

[13] highlights the benefits of GIS for enhancing data management and analysis, including the integration of remote sensing data. The limitations and challenges associated with standalone computer systems for air pollutant distribution simulation are briefly mentioned. Overall, this paper provides a valuable resource for those interested in using GIS to model air pollution in urban areas.

[4] Visualization developers user elementary graphical units called visual encoding to map data to graphical representation. Google Earth Engine built an interactive app (still at an experimental phase) that allows view changes in pollutant concentration (e.g. NO2) over time (last 30 days) using side-by-side or swipe map visualization and time series charts named TROPOMI Explorer.

[2] The World Air Quality Index (WAQI) project provides a map that shows last 48 hours data of PMs, O3, NO2 and SO2 as well as meteorological data. Some websites present air quality information, most of them specific to a country or a region [1] and static, with no interaction.

3 Research Plan

Environmental pollution is a significant issue that affects people's health and the planet's well-being. It is crucial to understand the severity and extent of pollution at a local level to address this problem efficiently. In this data visualization project, we explored the extent of pollution in different counties across the United States and highlight the need to take action against it. Environmental pollution is a serious problem that is affecting millions of people and our planet at the same time. It is really important to facilitate people to understand the severity of this problem. Through this project, we aimed to show the extent of pollution in different counties and how social and economic factors play a role in them, allowing users to understand the need to reduce pollution. We gathered data from the various sources which could be directly linked to or in a way be the cause of pollution or might be contributing factors to it from the sources mentioned in Section 1. The difficult parts that we thought for this project was to collect the data from various sources, pertaining to the counties, and finding a common field in the different datasets that would allow us to combine them into a single file. Another challenge which we faced was how would we handle if there is any missing data in any of the datasets, since we did not want to skew the conclusions or let the data we plan to visualize to be misinterpreted. We also found a way such that multidimensional data can be represented in a form which can allow the user to easily understand the different dimensions of the data and draw conclusions without any complex analysis.

We identified the major contributors to air pollution in the United States by analyzing two key sources of data - commute data and air quality datasets across different counties. Commute data, encompassing both public and private transportation, has been examined in terms of intra-county travel, inter-county travel, and cross-county commuting to understand the role of transportation in air pollution. The air quality dataset has been analyzed to determine the current Air Quality Index (AQI) for each county, as well as the levels of PM2.5 and PM10 emissions, which are significant pollutants. To contextualize the relationship between air pollution and socio-economic factors, we also incorporated the income and population data of the respective counties. By examining the data from multiple dimensions, the visualization elucidates the variation in air pollution levels across different counties in the United States. To effectively present this multidimensional data, the visualization employs several advanced data visualization techniques such as geo-map, scatter plots, and spider plots. The geo-map provides an intuitive and geospatial representation of the counties, allowing us to visualize the data related to it. The scatter plots are used to compare the various sources of pollution and analyze their interrelationships. The spider plot is employed to create a polygon that displays all the parameters, offering a comprehensive understanding of the pollution levels in a particular county.

3.1 Data

We are using a multitude of data from different sources. We are working with county-level data. To represent the counties, we have GeoJSON data representing the counties and their coordinates. We also collected population and income data from the Census Bureau and environmental data related to pollution from the Environmental Protection Agency. Within the pollution data collected from the EPA, we have NO2 levels, PM25 data, and Ozone Concentration data. These attributes are used to visualize the pollution data at the county level. We also collected commuting data from the Department of Transportation. This commuting data has 20 attributes. To begin, we standardized the format of all datasets to ensure consistency and avoid any data discrepancies. This is an important step as it allows for a seamless integration of data across multiple sources. Additionally, by adding a primary key to each dataset, we created a uniform identifier that can be used to merge different datasets and enable cross-referencing of data points.

Furthermore, we used Python scripting to combine all datasets into a single file, which is a GeoJSON file in our case, making it easy to

visualize and analyze the data. By aggregating the pollution data at the county level and using the county-name-state-name format as a primary key, we ensured that each data point is linked to a specific county.

The inclusion of commuting data is also a valuable addition to our dataset, as it provides insight into the transportation patterns within each county. By adding the commute that happens within the county and the commute that happens from people coming from different counties to the county, we accounted for the majority of commuting that happens in and out of each county.

Missing data is a prevalent issue in datasets, and it is imperative to handle it appropriately to ensure the integrity of the data and the accuracy of any analyses or models built on it. Our approach of using an adjacency list of counties surrounding the county to fill in the missing values is a valid method that has been employed in many studies to impute missing data. This approach is based on the assumption that neighboring counties are likely to have similar characteristics and that the missing data can be estimated by taking the average of the available data from those neighboring counties.

However, it is essential to consider the potential limitations of this approach, as the similarity of neighboring counties may not always hold, especially if there are significant differences in environmental, demographic, or socioeconomic factors. Moreover, the accuracy of the imputation can vary depending on the number and quality of neighboring counties used and the type of missing data, whether it is missing completely at random, missing at random, or missing not at random.

3.2 Visualizations

We used various visualizations in our project.

Spider plots, also referred to as radar charts or web charts, offer a comprehensive and visually appealing means of displaying multivariate data in a two-dimensional format. These charts are particularly useful in comparing multiple variables or data sets on a single graph, facilitating the identification of similarities and differences between the variables under consideration. Their compact format and ability to highlight trends or patterns in the data make spider plots a popular choice in various fields, including business, marketing, and sports, where they are used to analyzing and visualize performance data. In this study, we utilized spider plots to visualize pollution data, with the area covered by the points representing the pollution present in the county. Through the comparison of different variables, we aimed to identify the different kinds of pollution present in the area. While spider plots provide valuable insights into the presence and distribution of pollution, they are not a definitive means of identifying the reasons behind them.

Scatter plots offer a powerful means of displaying the relationship between two variables, and are particularly useful in cases where large amounts of data must be analyzed and presented in a visual format. In this study, we utilized two scatter plots to compare pollution attributes with two social attributes: income and population. As the population is directly proportional to pollution, while income is inversely proportional, these social attributes were chosen to provide insight into the relationship between pollution and social factors. Our analysis of the scatter plots revealed that counties with a high level of manufacturing activity tend to have higher levels of pollution, but lower levels of income. Conversely, counties with high populations tend to have lower levels of pollution, but higher levels of income. These findings provide valuable insights into the complex interplay between environmental factors and social dynamics, and highlight the need for careful consideration of these factors in policy decisions aimed at addressing environmental issues. Overall, scatter plots served as a powerful tool for visualizing and analyzing data, and offer valuable insights into the relationships between different variables.

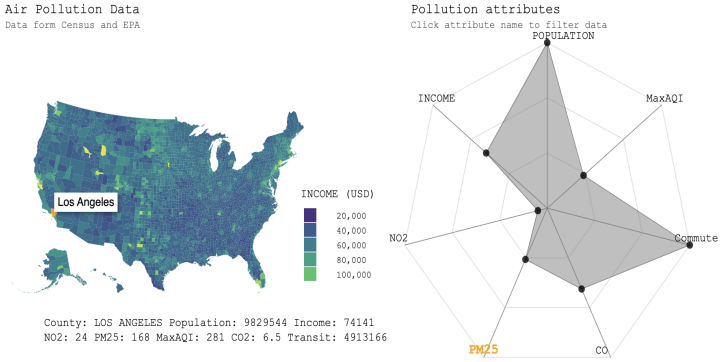
Geo maps are an effective means of visually displaying data tied to specific locations, such as demographic or environmental data. They can aid in identifying patterns and trends in the data, providing valuable insights into the distribution and concentration of the data across different regions. Geo maps are particularly useful in fields such as environmental science, public health, and urban planning. In our study, we leveraged the power of geo maps to represent income data across the counties. By displaying this data in a visual format, we were able to identify areas with high or low income levels, and determine the distribution of income across different regions. Overall, the power of geo maps in analyzing and visualizing spatial data cannot be overstated, and they offer a valuable tool for researchers and policymakers alike to better understand the distribution of data across different regions and make more informed decisions to address issues related to economic disparities.

A search box that allows the users to search for specific counties and add them to a table can be extremely useful for comparison because it makes it easier for users to find and isolate the counties they need. With a search box, users can quickly enter a county name or search term to locate the specific county they are looking for, without having to manually sort through a large amount of data. This can save significant time and effort, especially when working with many counties. Once the user has located the county they need, they can easily add it to a table for comparison with other counties. This can help to highlight similarities and differences between different counties and can provide valuable insights into the patterns and trends within the data.

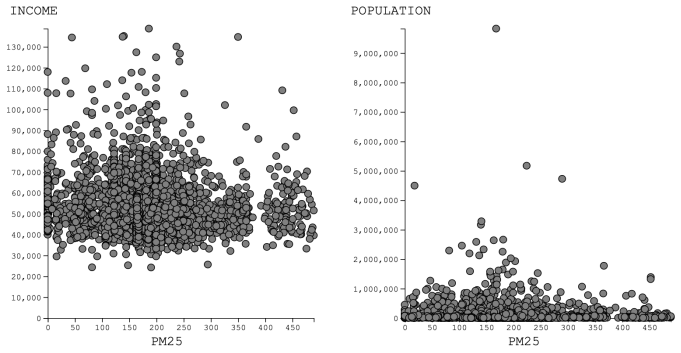
In this research, we utilized several technologies to create interactive data visualizations that enable users to explore and analyze complex data sets easily. By leveraging the power of D3.js, HTML, and CSS, we were able to create visualizations that enable users to explore the relationships between multiple variables across different dimensions. Our approach also utilized the properties of JSON format, which made it easy to separate and structure the required data for the visualizations. Moreover, we added interaction techniques within the visualizations, which augmented the visualization capabilities.

4 Results

We incorporated multiple interactive features into our dashboard, providing users with a seamless and intuitive experience. By hovering over a county in the geo-map, the spider plot for that county is displayed, providing users with a detailed comparison of different attributes.



The information displayed in the spider plot, including commute data, pollution levels, and population, enables users to compare and contrast counties across different dimensions. Additionally, by clicking on the spider plot attributes, the scatter plots display data for those attributes, allowing users to explore the relationship between two variables.



The scatter plot points represent individual counties, and hovering over them highlights the corresponding county in the geo-map, making

it easier to see the location of each county. The ability to easily compare and contrast counties based on multiple attributes, combined with intuitive visualizations and interactions, provides users with a powerful tool for exploring and analyzing complex data. There is also a search box which allows the users to compare the different counties and the data associated with that county, this makes it very easy for the users to compare different counties and their data. The search box has an autocomplete feature, which makes it easier for the users to find different counties and add them to the table view.

Compare Countries

Enter the name of the countries you want to compare

County Search Clear

County	Income	Population	MaxAQI	PM2.5	Commute	NO2	CO
Los Angeles	74141	9829544	281	168	4913166	24	6
New York	70947	1576876	154	157	973683	28	2
San Francisco	135189	815201	84	140	594574	13	6
Madison	59852	395211	105	225	173427	13	0
Lee	45002	177218	97	120	79150	13	0
Autauga	48347	59095	46	193	39703	13	0

The evaluation of our dashboard was guided by several key principles, including clear communication of purpose, data accuracy, interactivity and exploration, accessibility and inclusivity, and performance. Our dashboard was designed to present pollution data in the different counties in the United States with an emphasis on socio-economic factors like income and population. This will allow the users to explore the pollution data and gain insights in to the factors contributing to it.

To ensure clear communication of purpose, the dashboard had clear titles, labels for each visualization, which made it easy for users to understand the data being presented and draw conclusions. Additionally, we provided a brief description of the data and its sources used to collect the data, enabling users to understand the provenance and validity of the data.

Data accuracy was also a critical consideration in the evaluation of our dashboard. We carefully evaluated the reliability of our data sources, identified any anomalies or errors, and ensured that the data was presented in a clear and transparent way. In case of missing errors, we used the method to fill the data using the data from surrounding counties. Any anomalies or errors in the data were identified and addressed promptly to ensure data accuracy.

Interactivity and exploration capabilities were also essential components of our dashboard. We provided users with the ability to select the data from different counties, compare the data associated with different counties together, change the different attributes in scatter plots to understand the trends using the scatter plots and explore the data further to gain deeper insights. This level of interactivity enabled users to draw conclusions from the data that may not have been immediately apparent from the dashboard alone.

To ensure accessibility and inclusivity, we used clear labeling to make our dashboard easy to read and understand for a wide range of users. We also tested the dashboard's performance on different devices and browsers to ensure that it could be used effectively by a diverse range of individuals. The dashboard was easy to read and navigate, providing a smooth user experience.

For the evaluation of the performance of our dashboard, the dashboard was evaluated for its loading speed, smoothness of performance, and ease of use. The dashboard upon initial load takes a few seconds to load, but once it is loaded the performance was smooth, even with a large geo-json file and the attributes it contains like ours. Navigation was intuitive, allowing users to explore the information that was provided in the dashboard quickly.

5 Conclusion

We aimed to develop a dashboard that provides users with a comprehensive view of air pollution levels, socio-economic factors, and commute data in the United States. We analyzed and processed a vast amount of data to create the visualizations and implemented interactivity to enable users to explore the data further. Through this project, we have learned several important lessons. First, we recognize the importance of data accuracy and transparency in creating effective visualizations. Without reliable data sources and clear explanations of the data presented, the conclusions drawn from visualizations may be inaccurate or misleading. Therefore, we made sure to carefully evaluate the quality of our data and provide users with transparent and detailed information about its provenance and validity. We also accounted for a way to tackle the missing data, this method was an attempt to provide the data in a non-biased way to the users.

We also learned that interactivity and exploration capabilities are essential for creating a successful visualization. By allowing users to manipulate and filter data, explore specific areas of interest, and draw insights that may not be immediately apparent from the dashboard alone, we were able to create a dynamic and engaging experience for users. For future work, there are several directions in which this work could be expanded upon. For example, we could incorporate additional data sources to provide a more comprehensive picture of environmental and socio-economic factors at the county level. Additionally, we could explore ways to incorporate machine learning and predictive modeling techniques to forecast future trends in pollutant levels and their impacts on human health. We can even expand this to work to include all the regions in the world if a dataset for that is available. We can use a time series way of visualizing the data for each year allowing for the users to go back and see how the pollution levels have changed over time, giving them a way to better understand the pollution level in their county. We thought of making the dashboard accessible to the color-blind by using colorblind palate but couldn't achieve that, so allowing the users to choose their own colors could be a great addition, displaying the color based on the attribute by the user can also be added to this project.In conclusion, we believe that our work in creating an interactive dashboard has the potential to promote greater awareness about the impact of environmental factors on human health.

References

[1] Airparif, "air quality forecast.". `https://www.airparif.asso.fr/en/#`.

[2] World air quality index project, "world air quality index.". link : `http://my.url.com/`.

[3] M. Bostock, V. Ogievetsky, and J. Heer. D-3: Data-driven documents. *IEEE transactions on visualization and computer graphics*, 17:2301–9, 12 2011. doi: 10.1109/TVCG.2011.185

[4] J. Braaten. Monitoring air quality with s5p tropomi data. 4 2020.

[5] S. Dinda. Environmental kuznets curve hypothesis: A survey. 12 2004.

[6] A. Häägg, C. Weil, and N. Rönnberg. On the usefulness of map-based dashboards for decision making, 07 2020. doi: 10.36227/techrxiv.12738683.v1

[7] N. Janssen, P. Fischer, M. Marra, C. Ameling, and F. Cassee. Short-term effects of pm2.5, pm10 and pm2.5-10 on daily mortality in the netherlands. *The Science of the total environment*, 463-464C:20–26, 06 2013. doi: 10.1016/j.scitotenv.2013.05.062

[8] C. Jing, M. Du, S. Li, and S. Liu. Geospatial dashboards for monitoring smart city performance. *Sustainability*, 11:5648, 10 2019. doi: 10.3390/su11205648

[9] P. Jonnalagadda, C. M. Swoboda, H. G. Priti Singh, S. Scarborough, I. Dunn, N. Doogan, and N. Fareed. Communicating area-level social determinants of health information: The ohio children's opportunity index dashboards. 2021.

[10] M.-J. Kraak. Geovisualization illustrated. *ISPRS Journal of Photogrammetry and Remote Sensing*, 57:390–399, 04 2003. doi: 10.1016/S0924-2716(02)00167-3

[11] H. Li, H. Fan, and F. Mao. A visualization approach to air pollution data exploration—a case study of air quality index (pm2.5) in beijing, china. *Atmosphere*, 7:35, 02 2016. doi: 10.3390/atmos7030035

[12] J. Li, S. Chen, K. Zhang, G. Andrienko, and N. Andrienko. Cope: Interactive exploration of co-occurrence patterns in spatial time series. *IEEE Transactions on Visualization and Computer Graphics*, 25(8):2554–2567, 2019. doi: 10.1109/TVCG.2018.2851227

[13] L. Matejicek. Spatial modelling of air pollution in urban areas with gis: a case study on integrated database development. *Advances in Geosciences*, 4:63–68, 2005. doi: 10.5194/adgeo-4-63-2005

[14] L. Matějíček, P. Engst, and Z. Jaňour. A gis-based approach to spatio-temporal analysis of environmental pollution in urban areas: A case study of prague's environment extended by lidar data. *Ecological Modelling*, 199:261–277, 12 2006. doi: 10.1016/j.ecolmodel.2006.05.018

[15] L. Matějíček, P. Engst, and Z. Jaňour. A gis-based approach to spatio-temporal analysis of environmental pollution in urban areas: A case study of prague's environment extended by lidar data. *Ecological Modelling*, 199(3):261–277, 2006. Ecological Models as Decision Tools in the 21st Century. doi: 10.1016/j.ecolmodel.2006.05.018

[16] W. Shi, M. Wong, J. Wang, and Y. Zhao. Analysis of airborne particulate matter (pm2.5) over hong kong using remote sensing and gis. *Sensors (Basel, Switzerland)*, 12:6825–36, 12 2012. doi: 10.3390/s120606825

[17] R. Taborda, N. Datia, M. Pato, and J. M. Pires. Exploring air quality using a multiple spatial resolution dashboard — a case study in lisbon. In *2020 24th International Conference Information Visualisation (IV)*, pp. 140–145, 2020. doi: 10.1109/IV51561.2020.00032

[18] Y. Zheng, F. Liu, and H.-P. Hsieh. U-air: When urban air quality inference meets big data. pp. 1436–1444, 08 2013. doi: 10.1145/2487575.2488188

[19] C. Zuo, L. Ding, and L. Meng. A feasibility study of map-based dashboard for spatiotemporal knowledge acquisition and analysis. *ISPRS International Journal of Geo-Information*, 9:636, 10 2020. doi: 10.3390/ijgi9110636