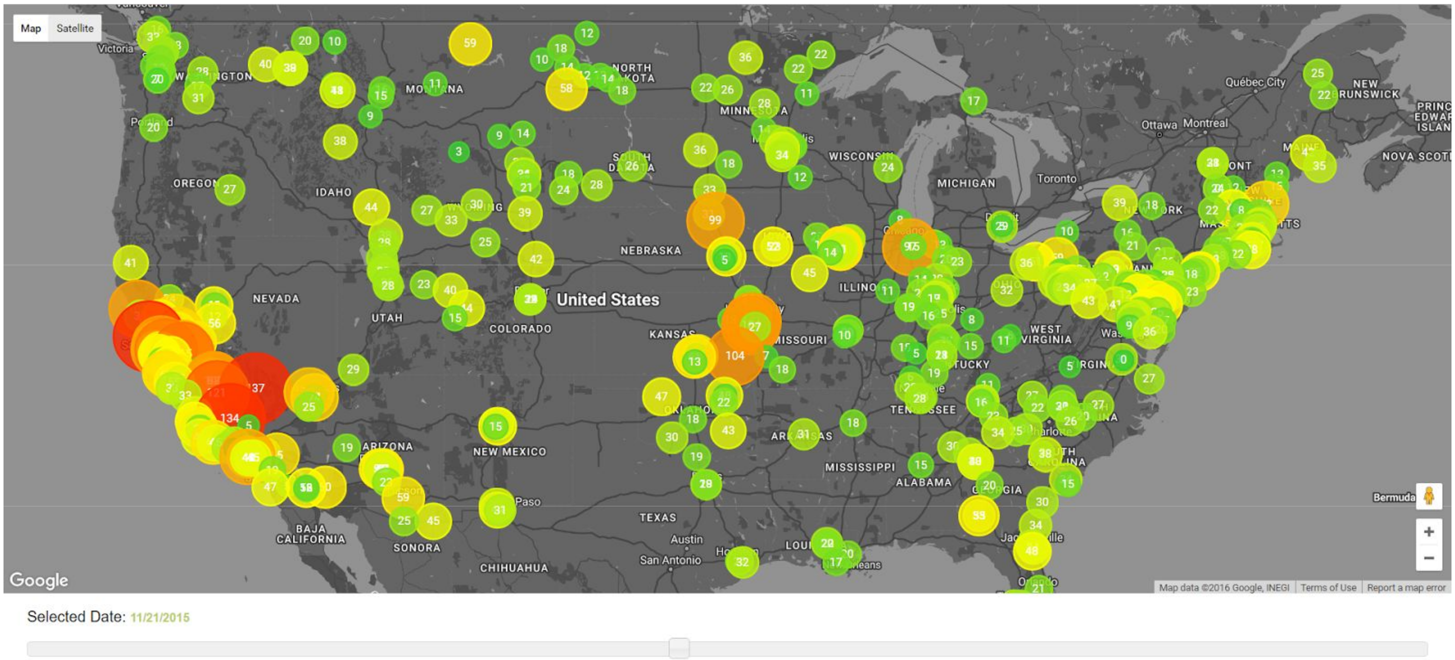
Time series Google Maps visualization of AQI for PM2.5 pollutant data and handling big data.

Ai-Linh Alten Brazen Abelgas

University of California, Merced



Selected Date: 11/21/2015

Figure 1**: Visualization example.** AQI from monitoring sites for given day 11/21/2015 all points converted to UTC. AQI determined by a daily average. AQI levels determined by color scale: Green = Level 1, Yellow = Level 2, Orange = Level 3, Red = Level 4, Purple = Level 5, Maroon = Level 6.

Abstract

The study demonstrates the use of ground-level data to map the trend of ambient air pollution, specifically fine particulate matter with aerodynamic diameter less than 2.5 micrometers ($PM_{2.5}$), for the purpose of quantifying and classifying its concentration over the United States over a period of time. This is done in whole to create a visual representation of the spread of air pollution and increase overall awareness of the problems that lie with it. Air pollution is a major hazard to both the environment and our health, having a direct correlation to climate change, depletion of the ozone layer, and respiratory and cardiovascular problems.

Although the United States has seen a decline in the concentration of particulate matter over the last few decades, largely in part due to the Clean Air Act, it is still a substantial issue that will continue to require constant care and attention.

The concentration of fine particulate matter is shown through a time-series visualization of data collected from the publically available ground observation data from the EPA combined with the use of Google Maps API. The study used measurements of ambient $PM_{2.5}$ from the database between the years 2010 and early 2016. Different levels of the air quality index (AQI) were differentiated by color in order to quickly identify separate hazardous regions.

Keywords: AQI, PM2.5, temporal map visualization, time series, big data

1 Introduction

Since preindustrial times, human activities have considerably increased the concentration of fine particulate matter in both urban and rural regions (Silva et al 2013). $PM_{2.5,}$ is a small but particularly important part of a larger picture known as air pollution, one of the larger threats to humanity. $PM_{2.5}$ is able to completely bypass innate human defenses and is able to deposit deep into the lungs and enter directly into the blood stream (National Research Council). Research has found that $PM_{2.5}$ is highly associated with negative health impacts such as premature mortality, respiratory and cardiovascular morbidity such as aggravation of asthma and respiratory symptoms, and mortality from cardiovascular and respiratory diseases (Voiland 2010; WHO – National Office of Europe 2013).

One of the driving forces behind the growth of fine particulate matter is the burning of fossil fuels. Fossil fuels are burned in order to supply vehicles with energy for our basic needs of transportation, and although hybrid cars are helping to reduce the amount of fossil fuels used, they still burn regular gasoline and only delay the problem. Fossil fuels are also used to make products such as medicine, cosmetics, plastics, etc., and they are used to provide electricity ("What are fossil fuels used for?" 2015). Over 65% of the electricity generated in the United States in 2015 was from fossil fuels alone (U.S. Energy Information Administration). Every year, more vehicles are produced, more factories and power plants are built and continue to pump out smog, and the global concentration of ambient air pollution will only continue to increase unless more necessary measures to further reduce pollutant emissions are quickly taken on a global scale.

Air quality in the US has significantly improved in the past several decades thanks to the Clean Air Act. It has allowed the EPA to establish National Ambient Air Quality Standards (NAAQS) in order to decrease air pollution within the country and protect public health. In order to meet those standards, both stationary and mobile sources were regulated in order to decrease the air pollutants emitted, which includes power plants, oil refineries, cars, and planes ("Summary of the Clean Air Act," 2016; "Sources of Air Pollution," 2013). Still, in the Los Angeles area alone, several thousands of people die annually due to respiratory diseases alone, which could have possibly been avoided if the pollution levels in the area were far lower than they currently are (Cromar et al. 2016). The amount of pollution-related deaths could rise, however, if United States soon becomes unable to make use of either the Clean Air Act or the EPA if a certain President-elect follows through with his plans to cut the EPA and pollution-related regulations in order to "save businesses money" (Jerde 2015).

Not all regions of the world have had the opportunity to benefit from effective protective environmental measures, especially in low- to middle-income countries, where it is needed most (“Air pollution and population density”). Globally, ambient air pollutions kills approximately 3 million people per year, and 87 per cent of these deaths are found in developing countries, whom are typically unable to properly establish countermeasures against air pollution (“Ambient Air Quality and Health,” 2016). It is also important to note that on average, pollution roughly triples per ten-fold increase in population of a city (Hansen 2016).

Air pollution is calculated typically through either land-based measuring techniques such as analyzing contaminants collected from a filter or canister, or through remote sensing techniques in which measurement tools such as satellites and planes measure the electromagnetic energy reflected from the earth (B.C. Air Quality; “What are fossil fuels used for?” 2015). Ground-level data for PM is scarce, so remote satellite sensing combined with modeling is often used to assess the population exposure at country-level scales (WHO – National Office of Europe 2013). The precision of the results is largely dependent on the availability of surface measurements, but it still serves as a relatively accurate tool for estimating PM over a large area. For the scope of this project, hourly data from numerous ground-level air quality monitors throughout the United States has been collected and merged with global annual average $PM_{2.5}$ grid data (in the form of GeoTiff) from the National Aeronautics and Space Administration (NASA)’s MODIS satellite to model the air quality index across the surface of the country. The satellite provides a measure of Aerosol Optical Depth (AOD), - the degree to which aerosols prevent the transmission of light by scattering or absorbing it throughout the entire atmospheric column.

Creating an accurate and precise representation of fine particulate matter for both a specific location and time period can be difficult, however. The data previously mentioned can be considered to be an extremely large data set, further increasing the complexity of the analysis. As the region analyzed is of the United States rather than the global scale given by the GeoTiff data, it had to be parsed in order to extract the data relevant only to the specific country, per year. The data then had to be mapped to the Google Maps API, which unless properly optimized, would create longer loading times than necessary if the architecture and infrastructure behind the database is unable to keep up with the client's requests. Any outliers to the data had to be properly addressed and handled as well.

The objective of this project is to visualize environmental hazards data of $PM_{2.5}$ throughout the United States for the purpose of increasing awareness of the issue that is air pollution, and subsequently be aware of the health and environmental issues related to it. The project will use measurements of ambient $PM_{2.5}$ from the EPA AQS database for the years 2010 to 2016 as well as the $PM_{2.5}$ estimates derived from MODIS AOD data.

2 Dataset

Dataset used for this project was pre-generated vector data from U.S. EPA. The dataset was a daily summary of AQI of the PM2.5 particulate. EPA offers data listed by year, in reverse order, back to 1990. Each file is updated twice per year: once in June and once in December. The dataset we visualized is from January 2010 up to June 2016.

3 Method

Our project goal is to spread awareness of pollution throughout the U.S. by creating a visual that those who visit the webpage can understand and interpret. For this project, we focused on designing the project in such a way that the user will be able to interact with the map.

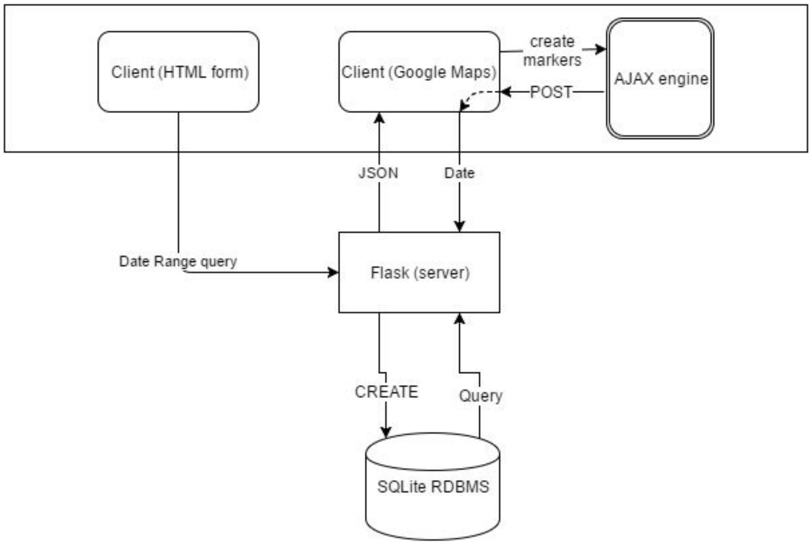


Figure 2: **3-tier architecture.** Based off client-server model. Diagram shows interaction between client, server, and database for this project. PM2.5 is stored in SQLite database and queried with AJAX.

The framework used to visualize the data and allow user interaction is a simple 3-tier architecture or client-server model (Figure 2). This architecture consists of the Client – Google Maps, Server - Flask, and Database – SQLite RDBMS.

3.1 The Client

The client (or browser) renders an HTML template from the Flask server to visualize the Google Map. The user can select a date to visualize more PM2.5 data points of the given day. All locations have their corresponding local dates converted to UTC to keep the time visualization concurrent. The browser interacts with the AJAX engine to submit POST requests to the server in order to return back a JSON object consisting of all the locations and the AQI of the given day.

3.2 Flask

Flask is a Python WSGI framework that acts like a server. The server does backend processing for the web application. This is the middle-man between the client and the relational database. For the project, we used the CSV module and Flask-SQLAlchemy to parse the pollution data CSV files and create the necessary table containing the site ID, local date, latitude, longitude, and AQI value. After database is created, the server is then ready to interact with the client and respond to its requests. When client makes an asynchronous request with AJAX, Flask looks up through database with a simple query that will return the JSON object with the necessary AQI values and locations.

3.3 SQLite RDBMS

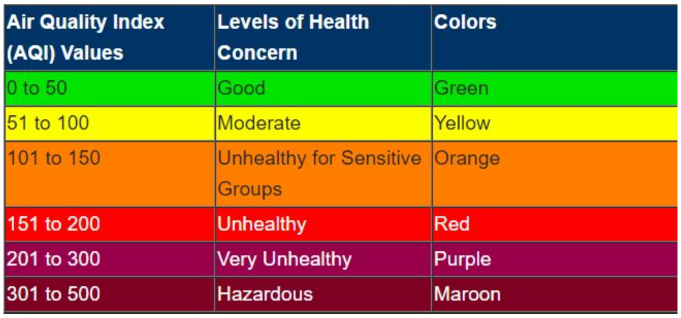
I purposely chose the SQLite RDBMS since it goes hand-in-hand with Flask's SQLAlchemy. Generally, when handling Big Data, having a table size that grows will be inefficient when trying to make faster query times. However, the necessary query we need to build the JSON object is still generally fast even with the bigger table size. When creating the schema, we made sure to index by date so that the index lookups will be roughly O(lg(n)). Currently, our database is a size of 114 MB with **n = 1,986,241** entries. The table consists of data between the dates January 1, 2010 to April 31, 2016.

4 Evaluation

To verify that the PM2.5 temporal map visualization displays relevant information, we performed two task-based studies: (1) Display daily AQI on map in such a way that humans can interact and can interpret it, and (2) be able to handle big data with a web application. Our studies use air quality indices reported from multiple EPA monitors all over the United States to demonstrate pollution levels in all parts of the country.

Materials This project will visualize a map from 2010 to 2016 with daily AQI based on PM2.5 pollutant from U.S. EPA's pollution monitoring stations across the United States. The vector data in CSV format includes longitude and latitude of each monitor and aggregated sub-daily measurements taken at each monitor. The single sample value the monitor takes is a daily sample (24-hour duration). The mean and max daily sample have the same value. There may be multiple records or missing records for the monitor if the pollutant is not sampled every hour, there are multiple pollutant standards, or if there was an exceptional event that monitoring agency may have had to be excluded from comparison to the standard.

4.1 The Google Map Visual



Air Quality Index (AQI) Values	Levels of Health Concern	Colors
0 to 50	Good	Green
51 to 100	Moderate	Yellow
101 to 150	Unhealthy for Sensitive Groups	Orange
151 to 200	Unhealthy	Red
201 to 300	Very Unhealthy	Purple
301 to 500	Hazardous	Maroon

Figure 3: **U.S. EPA AQI.** Air Quality Index that United States Environmental Protection Agency developed. This is divided into 6 levels that indicate health concern. This makes it easier for humans to interpret an index versus a pollutant's unit.

With Google Maps, we created a PM 2.5 concentration bubble map (Figure 1) represented by AQI values (Figure 3) for the given day. The data is from the U.S. EPA and the data represents sites that calculate the AQI daily based on PM2.5 concentration. The user can interact with the map slider that shows AQI values per day at every site across the United States. Over time, the bubbles will either scale up or down depending on the AQI. AQI is represented by gradient-based color scaling. AQI 0-50 are represented by a green to yellow color gradient, AQI 51-100 is represented by yellow to orange, AQI 101-150 is represented by orange to red, AQI 151-200 is red to purple, AQI 201-300 is purple to maroon, and AQI 301-500 is maroon to black. On days of known events, it can be seen when and where pollutants of high concentration are located. Such as September 12, 2015 around the time of the California Butte fire, the AQI was a high 320 at the San Andreas monitor.

4.2 Queries Execution Time

Year(s)	Size of DB	N (# rows)	Query time	Function calls	Response Time
2016	8.41 MB	77,865	0.051 s	5886	0.038 – 0.364 s
2015-2016	67.5 MB	467,955	0.170 s	21033	0.154 – 0.457 s
2014-2016	48.4 MB	837,431	0.264 s	21033	0.247 – 0.581 s
2010-2016	114 MB	1,986,241	0.549 s	11819	0.547 – 0.858 s

Figure 4: **SQL Query times.** Times that it took for SQLAlchemy to query and return a JSON object. Year(s) column represents dataset between the timespan. Increase in timespan shows to have increase size in database size and query time.

Above in Figure 4, the table shows the time it takes to execute a query and return a JSON object from Flask when a user requests a date. Increase in database size shows a considerable amount of time to execute the query. Response time to query and bring JSON to client also has some overhead. The varying response times depends on whether client is running another process that makes the response slower. In addition to longer execution times as the database grows, Google Maps takes about 400+ milliseconds to draw or update markers.

5 Discussion and Conclusion

Visual With Google Maps, it was nice to create visual of a daily average dataset to see changes in pollution level over a year. This made it easy to observe changes in season as well as find locations of certain events that caused high levels of AQI. There are other kinds of datasets that could help with visualizing PM2.5 concentrations, but EPA provided the best dataset for what we wanted to expect to find in our results. Originally, starting off with satellite GeoTiff data of PM2.5 values made it easy to perform other spatial analysis tools like zonal statistics, however, it was hard to do a time-series analysis. Since the raster data was continuous and PM2.5 values that were yearly averages, seasons and special events could not be observed very easily . But with the daily data, it worked best with generating these kinds of results.

Limitations Although the map successfully displays all monitoring stations in the United States, given that we are working with a large dataset, the map queries that draws the markers onto the map may not be done in a timely manner that the client can have the patience for. 0.5 seconds to draw a map doesn’t sound like a long time, but trying to query multiple days at once may take a while for results to come up. The web application still has the capability of handling human interaction, but yet the response time in order for the human to receive and interpret the information will take longer than they would like. Relational databases can still be a good database to use even knowing that an increase in table size may slow down the query time execution. From the results, the times between a smaller and a larger database did not make the difference in time vary too much. The only real problem may be the Google Maps API itself. If it takes about 400 or more milliseconds to draw all its map markers, it can show even the API cannot handle large amounts of data.

Conclusion Some solutions to handle the problem of handling big data can be as simple as optimizing the way a user queries for a date to bring up a map visual. For one thing, Flask-SQLAlchemy has the capability to cache a query to make table lookup much faster. This can make the query execution times much quicker. To solve the problem of Google Maps not being able to display lots of markers onto a map, one can try to aggregate the markers in a way that humans interacting with the map can understand. An example would be: combining the markers to create a pie chart when the user zooms out of the Google Map. When zooming in the markers would separate back into their original locations. Overall, our results on the map visual create a good sense of change in pollution concentration daily and still can observe changes in season or other natural disasters.

References

"Air Population and Population Density." *Population Matters* (n.d.): n. pag. Web. 4 Dec. 2016.

"Ambient Air Quality and Health." *World Health Organization*. World Health Organization, Sept. 2016. Web. 03 Dec. 2016.

“B.C. Air Quality." *How We Measure Air Quality*. BC Air Quality, n.d. Web. 02 Dec. 2016.

Cromar, Kevin R., Laura A. Gladson, Lars D. Perlmutt, Marya Ghazipura, and Gary W. Ewart. "Estimated Excess Morbidity and Mortality Caused by Air Pollution above American Thoracic 201. *American Thoracic Society*. Marron Institute of Urban Management, 8 May 2016. Web. 5 Dec. 2016.

Hansen, Kathryn. "NASA Scientists Relate Urban Population to Air Pollution." *NASA*. NASA, 19 Aug. 2013. Web. 03 Dec. 2016.

Jerde, Sara. "Trump Says He Will Cut The EPA As Prez: 'We'll Be Fine With The Environment'" *TPM*. TPM Media LLC, 18 Oct. 2015. Web. 04 Dec. 2016.

National Research Council, (US) Chemical Sciences Roundtable. "What Are Small Particles and Society." *American Thoracic Society and Marron Institute Report* 13.8 (2016): 1195-

Why Are They Important?" *Challenges in Characterizing Small Particles: Exploring Particles from the Nano- to Microscale: A Workshop Summary.* U.S. National Library of Medicine, 2012. Web. 03 Dec. 2016.

Silva, Raquel A., Jason West, Yuqiang Zhang, Susan C. Anenberg, Jean-Francois Lamarque, Drew T. Shindell, William J. Collins, Stig Dalsoren, Greg Faluvegi, and Gerd Folberth. "Global Premature Mortality Due to Anthropogenic Outdoor Air Pollution and the Contribution of past Climate Change." *IOPScience*. IOP Publishing Ltd, 11 July 2013. Web. 2 Dec. 2016.

"Sources of Air Pollution." *National Parks Service*. U.S. Department of the Interior, 10 Jan. 2013. Web. 04 Dec. 2016.

"Summary of the Clean Air Act." *EPA*. Environmental Protection Agency, 17 Oct. 2016. Web. 04 Dec. 2016.

"U.S. Energy Information Administration." *What Is U.S. Electricity Generation by Energy Source? - FAQ - U.S. Energy Information Administration (EIA)*. U.S. Department of Energy, n.d. Web. 02 Dec. 2016.

Voiland, Adam. "New Map Offers a Global View of Health-Sapping Air Pollution." *NASA*. NASA, 22 Sept. 2010. Web. 02 Dec. 2016.

“What are fossil fuels used for?" *New Mexico Oil and Gas Association*. New Mexico Oil and Gas Association, 2015. Web. 03 Dec. 2016

WHO – National Office for Europe. "Health Effects of Particulate Matter." *Health Effects of Ambient Air Pollution*(2000): 115-37. *World Health Organization - National Office for Europe*. Regional Office for Europe of the World Health Organization, 2013. Web. 3 Dec. 2016.