Pollution Data by Year and State in the United States

INFO/CS 3300: Project 2

Weston Forster (wtf25), Sam Hamburger (sh786), Sam Weintraub (saw294), Brian Zahka (bdz3)

**A. Description of the Work Done by Each Team Member**

Brian - Created functionality for generating graphs displaying the counties in a state upon clicking on the state. In doing so, worked with the data in javascript extensively to get it into a nice format for the graphs. Worked with Sam H to integrate the graphs into the UI.

Weston - Found all the data (in the form of .csv files) and processed the entirety of the dataset with d3, making it usable for the different visualizations contained in the page. Also calculated some important averages for displaying the heat map correctly.

Sam H. - Did all things UI and also contributed to other parts of the project. Made scales for each parameter, corresponding to the range they occurred in in the data. Fine tuned the gradients to display as nicely as possible on the map.

Sam W. - Worked on a feature (which we ultimately had to abandon because it proved to be harder than we thought) of displaying clouds (proportional to pollution quantity) over different areas. Also did extensive pre-processing on csv files to make the program load much faster.

**B. Description of the Data**

Our primary data sources, pollution data for each state from every other year from 1990 to 2016 (as well as 2015), came from <http://aqsdr1.epa.gov/aqsweb/aqstmp/airdata/download_files.html#Annual> . Pollution data is divided up by pollutant, with PM10, PM25, Ozone, SO2, NOx, and CO ppm tracked in each state for the given years.

We imported our datasets via CSV files (one for each year), but found that the large amount of data caused a notable lag in our model’s initialization. As a result, we examined the datasets in order to figure out the minimum number of columns we needed for each data point (which corresponded to the CSV rows), and deduced that 80% of the columns in each dataset were unnecessary for our model to adequately function. We went through and deleted these redundant columns in Microsoft Excel, and then re-ran our model.

We originally worked with data from every year starting with 2000, but after seeing only trivial differences between adjacent years and experiencing long loading times, we decided it unnecessary and inefficient to use the odd-numbered years (excluding 2015, which we decided was significant enough because it included important information about the last 3 years).

When using Excel, we had to be careful to preserve the data in our CSV files. Our County\_Code column needed to be three digits in order to work correctly in our visualization, and in some cases had leading zeroes that Excel would delete if we saved a normal updated csv file. As a result, we had to make sure to save the County\_Code column in a special format (“000”) in order to ensure that it would function correctly for each imported dataset.

**C. Description of the Mapping from Data to Visual Elements**

*Color*

In our visualization, we use color to identify the amount of pollution for the given pollutant in each state. Darker shades denote higher amounts of pollution, while lighter colors denote lower amounts.

We also use color to distinguish between different pollutants - our model’s US map changes colors based on the pollutant one chooses to look at. The colors are chosen to reflect each pollutant’s theme and personality.

Unfortunately, we were unable to obtain data for every state for every year for every pollutant; as a result, certain states are colored black in the applicable visualizations.

*Map*

We decided to go for a us.json file and used the D3.js geoAlbersUsa projection in order to project our map. Since we found that pollution can affect all parts of the United States and not just urban areas, we ruled out using a US map scaled for population.

Originally, we had planned on projecting our map by county, rather than by state. However, some counties in each state didn’t have data, and this caused our projections to have several black counties scattered all across the country- this disrupted the appeal of the visualization. Once we decided to project our map by state, we arrived at the pollution ppm values for each state by taking the average for each county in the state. While not every county had data available, we quickly figured out that enough of them were represented to get a feasible representation for each state.

For each state, the user can click on it and see a line graph of the change in ppm of the selected pollutant in that state from 1990 to 2016.

*Interactive Tools*

Our model includes an interactive menu that allows users to toggle through pollutant options, and users can play around with this to look at data for each pollutant. Users can also toggle the slider at the bottom to see data from each available year.

**D. The Story**

Our project visualizes pollution in the United States as its effects differ by state and year. Graphs that appear when a user clicks on each state show the change in pollution over time for the given pollutant. We chose this topic because we wanted to examine the change in annual pollution over the past few decades, both regionally and in terms of magnitude, and we purposely chose to work with an assortment of pollutants that had prominent effects on our environment and well-being.

As evidenced by our state graphs and U.S. map visualizations, pollution as a whole has decreased over the last few decades for most pollutants. However, there are some exceptions to this rule, and there has been a noticeable regional shift of certain pollutants.

From 2000 to 2014, the levels of PM2.5 in the southeast, and later the midwest, decreased significantly compared to other states. The worst ozone pollution centered around southern states like Tennessee and North Carolina in the 1990s and early 2000s, but this seemed to dissipate by 2014; mountain states like Colorado and Utah exhibit significant ozone levels in all available years, which likely corresponds to their notably high elevations.

The story of the SO2 data can be attributed to the decline of the coal industry in West Virginia. From 1990 to 2008, the SO2 levels in West Virginia were notably higher than in any other state; however, they began to level off in the late 1990s and early 2000s and finally reached the same levels as other states in 2010. Coal production is known to emit SO2, but double digit declines over the past several years significantly cut the amount of this pollutant in this state. High SO2 levels are also apparent for certain years in Hawaii; this can be attributed to the smog often associated with the state’s volcanoes.

NOx levels are consistently high in urban states, such as the northeast, Illinois, and California, and mountainous ones, like Arizona, Utah, Nevada, and Idaho. However, the degree of th