Process book
Air condition of major cities in California

Names: Spencer Fronberg, Qian Zuo, Mattia Grespan

November 10, 2018

1 Overview and Motivation:

Many environmental factors can affect our health, but the one that has the most impact to us is air pollution. We breathe air to live and what we breathe has a direct impact on our health. Environmental Temperature and humidity are also important factors that influence human's living standard.

In the past, research focused on only one issue of air pollution, temperature and humidity. However, these three issues influence human's living standards a lot. We want to create a visualization combining these issues together that informs the user about the relationship between them in terms of the air quality. Moreover, we want to provide recommendations in decision-making in terms of air condition quality.

In this project, we will collect and show data of these three aspects for major cities in California. We chose California because is the most populous state in the United States and the third largest by area. Due to the fact that the state extends from the very south of the west coast up to the north region of the west coast, the climate ranges from polar to subtropical. These characteristics should guarantee considerable variations in the data for temperature, humidity and pollution across the state. Hence, we should get more noticeable results.

Additionally, California is the biggest hub of software engineering in the world. Therefore, we consider pertinent to give our colleagues some suggestions about such important aspects for human's health like fresh air, convenient temperature, and suitable humidity.

2 Related Work:

The following are links to some of the material that inspired us for this project:

1. Air Pollution: Invisible Killer. BBC World Services. The Real Story. Audrey de Nazelle - Imperial College London and Maria Neira - World Health Organization.

   This is a very interesting episode of a radio show from the BBC where they discuss in a clear way about the awareness of the risks posed by polluted air.

2. Research + Remote Pollution Sensor Telemetry.

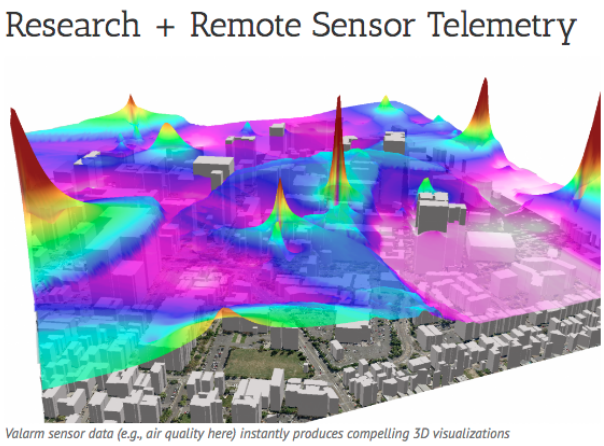

Figure 1: Valarm's company sensor data (Pollution)

This is a visualization of the air quality produced by sensor of the monitor industrial sensors Valarm. We found it interesting by the way it uses 3D encoding. This inspired us with a prototype version of our visualization.

https://www.valarm.net/research-and-academia/

3. Infoplease. Effects of Dry Air on the Body. https://www.infoplease.com/science-health/weather/effects-dr y-air-body

   We wanted to know more about the impact of lack of humidity on the body. We found this concise and well explained article.

4. We are considering different ways to encode friendly recommendations to the user about the levels of Pollution, Temperature, and Humidity. We found some good ideas like the following on Public Lab website: https://publiclab.org/notes/jiteovien/08-01-2018/air-quality-data-visualization-no-coding-necessary

Figure 2: Categorization of the Pollution levels

5. This example of a field shown during the Dataset Types lecture was also a good reference for us.

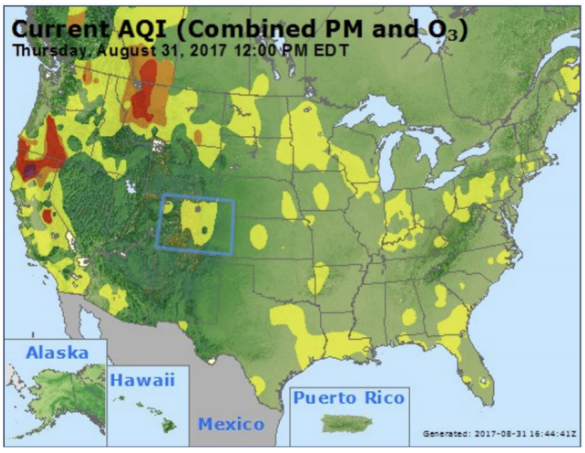

Figure 3: Air Quality map seen in class

6. One the most inspiring visualization we found online for our project was from the Air Pollution Quality Monitoring website *PurpleAir.com*. This website is key for our project because we also extracted the data from it. We would like to reproduce some of the visual encoding they use not only for pollution but also for temperature and humidity.

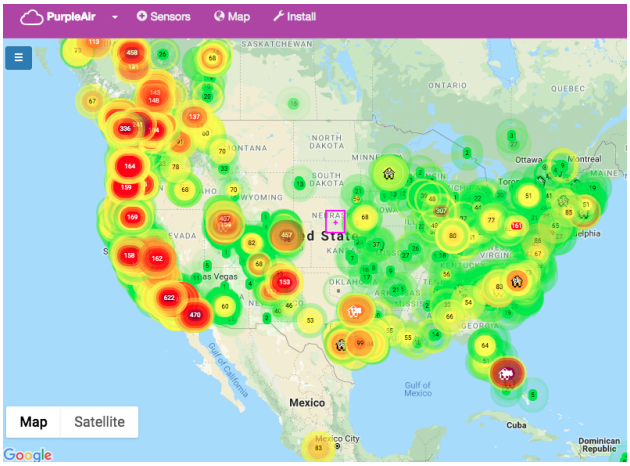

Figure 4: Purple Air visualization of North America Air Pollution Quality

3 Questions

- The first challenge/question we encounter at the beginning of this project was how to create a very friendly and informative visualization about the air conditions (temperature, humidity, and pollution). This visualization should help users to have a general idea about the air conditions in these cities in the simplest and interactive way possible.

- Were to find the right data about air quality condition?.
- Is there a noticeable relationship between pollution and temperature? What about pollution and humidity?.
- What would be the best time of the year to study the variation of these aspects of the air quality condition?.
- If we have to restrict our study to one state in the U.S. Which one would give us the best data to study these correlations?
- How to encode these tree aspects of air condition quality following the principles of design studied in class.

4 Data

We are downloading our data from the PurpleAir website (https://www.purpleair.com/sensorlist). Our data includes three different levels of pollution, but we are only focused on the PM2.5 particles level measured as CF ATM ug/m3. PM2.5 is a microscopic particle2.5 microns in width and almost 30 times smaller than the diameter of a human hair. When levels are high, PM2.5 particles form a haze in the sky, making their way into peoples respiratory tracts and reaching the lungs [3].

Our data also includes humidity(%) and temperature(F^o). Each entry in the data sets is recorded by sensors every one to two minutes. There are many different data sets that are available to us. We will focus on the vicinity of downtown (including it) areas for the major cities in California. There are about 20 to 30 sensors for each city. Our data is the format of csv files.

5 Data Processing

1. We had three factors to consider in order to find the best set of cities in California for this project.
   (a) The cities had to be well distributed all across the state. Ideally, from north to south and varying from big cities to rural cities in order to obtain a good variation in the values of the weather and pollution variables studied.
   (b) Many cities in California do not have enough sensors to get an accurate average of the air condition in them. This is why we needed cities with at least 20 sensor working properly and constantly during the time frame our study is focused.
   (c) Not every city in California has an available representation in geo.json maps for d3. This is why we had to choose cities that we could accurately show in our state map.

Under this constraints the optimal set of cities we found was: Sacramento, Eureka, San Francisco, San Jose, Bakersfield, Fresno, Los Angeles, and San Diego.

2. For each city, for each sensor data-set (between 20 and 30 csv files), we processed it to get the averaged pollution, humidity, and temperature in increments of 10 minutes. Obtaining new 10-minutes averaged sensor data-sets (csv) corresponding to every sensor.

3. For every city, we created a new data-set with the averaged pollution, humidity, and temperature from all the corresponding 10-minutes averaged sensor data-sets obtained in the previous step. Obtaining eighth final data-sets containing the averaged value for each variable in increments of 10 minutes for each city. These are the data-sets we are using for our visualization.

6 Design Evolution

Figure 5 and figure 7 show the digital sketch for one of the first alternative prototype designs we created.

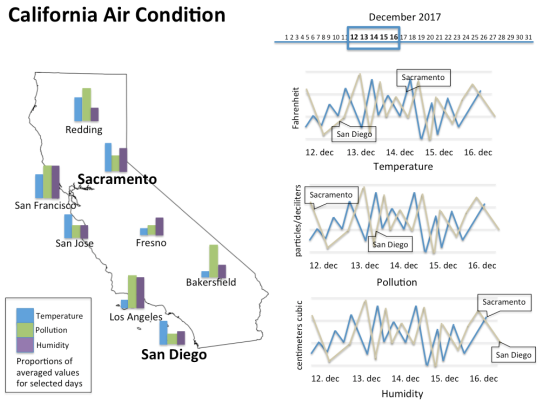

Figure 5: Prototype design

Some of the problems we found on this first approach were:

- Occlusion in case of adding more cities to the visualization. Furthermore, the little bar charts and the name of the cities would not be readable.

- Different units. Even though the intention of the bar chart is to show simple proportions of each of the values of the air conditions (pollution in blue, humidity in green and temperature in purple), the units of each of them are different. This is not a very good way to encode this values at once. We tried to encode this three values in different ways but none of them convinced us (See sketches in appendix).

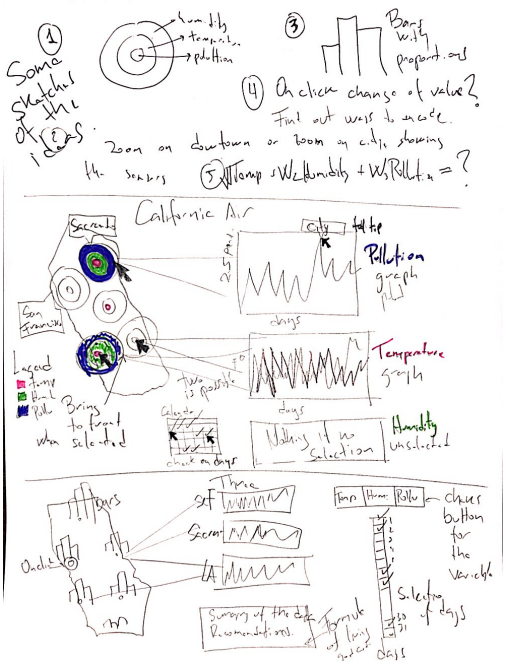

Figure 6: Prototype design

- In general one of the main problem we faced was to encode the three variables Temperature, Humidity and Pollution on the same map. We tried several options. Some of out different initial approaches to solve this can be traced in the sketches of the appendix.

The visualization we considered final on our project proposal is the following:
The decisions we propose as the final design for this visualization are the following:

1. Three buttons to select the condition of the air that the visualization will work with.

2. A brush bar to select any sequence of days. The visualization will use these information and generate the average of the values during that period of time.

3. Use the shape of the California state as an spatial region. This could be seen as a mark, but is also a channel to convey to give the user a better general sense of the location and distance between the cities the visualization is showing.

4. Circles to indicate each city (marks). This circles will be colored according to the value of the condition the visualization is currently showing. There is a color saturation bar with a specific hue gradient for each condition. Even though the saturation of color can be effective in showing the intensity of the selected air condition, we decided that the circles will also include the number of units that the color is encoding. Note that this last step is redundant but we think is important in order to the lecture of the data even easier to the user.

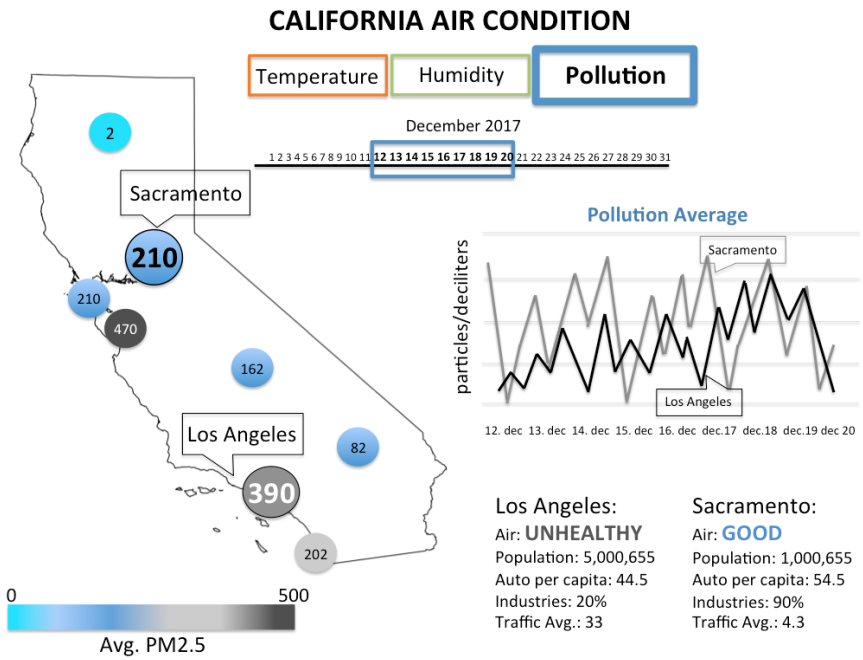

Figure 7: Proposal final design

5. Once the user interactively clicks (or hover) on a city, a plot of the time in function of the current selected air condition will show up. This plot will use time by hours. In order to compare the selected air condition on different cities the user can click more than one city and the graph for both cities will show up on top of the other.

6. Display more information of the city: population, number of autos, traffic, etc., as well as the final description summarizing the data for the current selected condition according to the data during the selected period of time.

7 Peer review and suggested changes

We received suggestions from our peer students Maitrey Mehta, Sai Varun, and Shalin Parikh.

1. On the first proposal for our visualization, we were going to use the average data of the sensor on every city per day for every data variable: Pollution, Temperature, Humidity. Our students peers found a problem on this approach and out intention of having a line chart plotting this data.

   If the user select only one day, the plot would consist on only one point. In general, for a small number of days selected, the plot was not going to be very helpful for purpose of the visualization.

They suggested us to take the average of the sensor every hour or every 10 minutes. This would make our plot chart more meaningful and will show more precisely the behaviour of the variables.

2. Our line charts are going to show only the information of two selected cities, since we think that if the data from more than three cities is displayed we may deal with undesired occlusion. However, they suggested us to try to include at least one more city on the selection to encompass more data insight for the user.

3. Another suggestion was us to change the bar gradient color for the pollution visualization. Our choice was from light blue to black. They think this choice of colors was counter intuitive.

4. They liked out prototype visualization 5 so we also discussed ways to fixed the problem we encountered with it. They suggested us to use an area chart, similar to the "Bill Gastes" one used in class for design critique, to represent the proportion of each variable on top of every city.

8 Implementation status

All of the following implementations are still in progress:

1. California State and cities map using Geo.json data.

2. Circles over the cities linked to the data: We have to figure out the best way to select and unselect cities for the comparison processes.

3. Line charts interacting with the data of the city selected: We are having some problem with the loading time of these charts. This has to do with the fact that we are showing a big amount of data (every 10 minutes signal).

4. Code for the city information area with suggestions for the user.

References

[1] Anton E. Kunst Casper W. N. Looman Johan P. Mackenbach. Outdoor Air Temperature and Mortality in the Netherlands: A Time-Series Analysis. American Journal of Epidemiology, Volume 137, Issue 3, 1 February 1993, Pages 331341, https://doi.org/10.1093/oxfordjournals.aje.a116680

[2] Infoplease. Effects of Dry Air on the Body. https://www.infoplease.com/science-health/weather/effects-dry-air-body

[3] The World Bank. IBRD.
Understanding Air Pollution and the Way It Is Measured. http://www.worldbank.org/en/news/feature/2015/07/14/understanding-air-pollution-and-the-way-it-is-measured