Project Goals

The primary objective of this project is to analyze the pollution patterns observed in a region that has a high degree of variability with regards to many factors. California makes an interesting case study because it's large, has areas with extremely high and low population, a wide range of climates, various biomes, and regions with vastly different altitudes. It's my hope that the large amount of variation between environmental conditions allows for many interesting observations and implications based on the types and intensity of aerial pollution present.

That being said, I think that the intended audience for this project is climatologists or other environmental scientists looking to analyze the patterns and vectors of air pollution. As such, I would like to reveal patterns and potential correlations between environmental/demographic variables and the severity of air pollution in the California area. My hope is that the relative importance of different variables will be revealed when considering the severity of pollution in an area, allowing for climate action to be taken with the focus on the most impactful variables. For example: in an urban area in the US with high population density, the biggest contributor to air pollution is almost certainly going to be emissions from vehicles. This means that a blanket policy that caps vehicle pollution limits won't have much of an effect on a rural town without many cars or people. This project aims to identify key pollution factors in different areas such that climate action can be focused on factors of high intensity. The intent of doing this via geovisualization is to present the data in a way that's accessible to the public and showcases spatial patterns such that anyone can look at the map and identify patterns.

Project Data

The key dataset that I used was the CalEnviroScreen 4.0 dataset. This data is derived from a government agency that regularly collects information on environmental pollutants and potential hazards. It includes a swathe of demographic data as well as a range of different environmental pollution levels. The dataset gets updated every year, but some of the sources remain a little bit ambiguous. Although there's a data dictionary present in the zip file that can be downloaded from the site, it mostly functions to explain what each attribute in the csv file represents - it doesn't explain the sources of the data very well. The data is collected at the census tract level, but it's not clear how often some of the data is collected. The environmental and toxic release inventory data is collected annually, but it's unclear how often the demographic data is collected. I believe that it's been derived from the US census website, which could lead to some degrees of inaccuracy if the demographic data is from a different year than the environmental information. What's more is that the data relies on collection sites for its environmental data, meaning that no matter what, there's going to be a certain degree of smoothing/averaging when mapping areas with choropleth techniques where areas are assigned an average or cumulative value.

This is apparent in the graduated symbols maps presented, where you can clearly see that data points are highly clustered in the center of the state and near the coast in the middle and Southern regions of the state. This is likely due to differences in population density of various areas, the government agency is more concerned with environmental hazards with relation to human health, making it much more pertinent to gather data in more populous areas. This dataset encompasses all of California and has decent coverage of the state, but there is a good deal of imbalance in terms of the distribution of datapoints. Since the datapoints are so much denser around areas with higher population, there are significantly fewer observations in the North and East regions of the state where there are mountains, deserts, and forests.

Maps and Objectives

In this project, I used graduated symbols maps to portray the PM2.5 particulate matter recorded in the air at each datapoint, as well as the total population within the census tract that each datapoint was recorded in. PM2.5 concentration is typically measured in micrograms per cubic meter of air. I felt that this method of mapping was appropriate for a recording of air pollution because of the way that aerial pollution spreads and disperses. The graduated symbols helps to convey the magnitude of pollution when it's observed at high values as it spreads through the air. The second map displaying the population total at each observation point was made because total population certainly has some correlation to the air pollution found in an area as people consume products that release emissions like gasoline for cars, electricity, and in-home HVAC systems. Additionally, the urban heat island effect increases with the amount of people living in a city as the city grows to accommodate the people and these are all variables I thought likely to correlate to air pollution. In short, I thought that a graduated symbols map would be a visually interesting way to portray the broad areas on the map that have high concentrations of PM2.5 and people while also revealing the bias in the spread of the data to maintain transparency with the audience.

The other visualization technique I used was a series of linked choropleth maps. I thought that this was an effective and time efficient way of displaying the patterns of multiple variables since I wanted to look at more than just two variables. The ones I chose to display were the PM2.5 concentration, the traffic density measured in “vehicle-kilometers per hour per road length,” poverty rate, and the rate of asthma emergency department visits. These choropleth maps are also accompanied by histograms that show the distribution of the datapoints across the range of values that they embody. A big thing to note with these maps is that the area of interest was shifted from all of California to just Los Angeles County. Paring down the area of interest allows for much easier interpretation of patterns than if the data were to encompass all of California. This also helps to alleviate some of the bias in the data distribution by limiting the area to Los Angeles County, which is a highly visible area with high population and has plenty of datapoints in it.

Observations/Future Work

The primary goal of the project was to investigate the relationships between variables and to identify potential patterns of pollution with relation to correlated variables like demographics and traffic. The hope is that this type of information would be useful for climate and environmental scientists when considering new policies or initiatives. Having information on which areas have the highest pollution concentrations and knowing what demographic factors might be contributing to that pollution level would help make policies more informed and effective.

I already spoke somewhat about the patterns seen the graduated symbols maps, but to reiterate, there was primarily a lot of high-value clusters near/inside Los Angeles County as well as in the center of the state along the mountain range. This led me to believe that not only are traffic and population density likely to be correlated with higher PM2.5 concentrations, but there could be a rain shadow effect happening along the mountain ranges. This could look something like: pollution from car exhaust is released into the air, wind from the sea blows all the air pollution inland, but the particulate matter in the air gets stuck on the mountain range and doesn’t quite make it over like the rest of the air does, which then might cause a buildup of PM2.5 along the mountains. This would require extensive field-work and further studies to prove though, but it would be an interesting environmental study to perform.

The highest pollution values in Los Angeles County were largely concentrated in the Southern half of the county, which checks out given that the Southern half is the hyper-urbanized, highly populated region of the county where there’s a lot of buildings, traffic, and emissions overall. The traffic map clearly shows the outlines of highways and major streets as being the most traffic-dense, with a lot of the clustering being in the Southern half as well as in the Western parts of the county moving up along the coast of California. The poverty map shows a lot of clustering in the Southern region of the county as well, with a lot of spots in the Northeastern corner of the county. I suspect that the northeastern poverty cluster being close to the airport has something to do with it since the loud noises of planes constantly landing/taking off probably drives rent prices lower and such. The asthma data is also interesting in that apart from the clusters in the urban center in the South (just like all 3 other variables), this variable also has a high rate of asthma in the Northeastern corner of the county like the poverty rate does. Every variable documented on the page showed some degree of clustering in the South-Central region of the county where the city is the most dense and urbanized, which was somewhat surprising to me.

The biggest limitation my data had was probably the form in which it was recorded. It was recorded at a very fine scale, meaning that it was hard to display the data on a county level. The sheer number of datapoints paired with the disproportionate spread of them also made this data really hard to interpret on a large scale. As such, I limited the scale on my second visualization method to the LA County region, which was much simpler to analyze. The graduated symbols maps were pretty successful at showcasing broad trends with high-pollution areas being clustered up, but it became hard to see some of the points with all of the overlapping symbols, even with the opacity of the symbols turned down somewhat. For the linked choropleth maps, I think that the biggest weakness was simply that the variables measured weren’t all in the same scale. This means that although you can see what spots have the highest or lowest values for each variable, it’s hard to discern what each of those values is. The dark red on one map is a completely different range of values from the dark red on the other maps. There can be some limited spatial pattern interpretation here, but the most helpful thing would be a statistical data analysis of some kind to measure p values and find statistical correlation between variables.

That would be the first thing I do in future works, would be to do statistical tests on a plethora of the variables to find out which variables are correlated to one another. It may be less accessible than a map, but presenting statistical test results would provide much more concrete evidence to prove correlation than looking at a pair of maps measured on different scales. I also think some bivariate choropleth maps could be useful here, but I decided to try out the linked choropleth maps instead in this case. The bivariate maps paired with statistical tests and p values would make for both strong evidence of correlation as well as easily accessible/interpretable maps to act as visual aids to make the project more credible in the future.