Logo, company name

Description automatically generated

Final Project

Water Quality Versus Blood Lead Levels

Location Analytics CIS3350.01

Dr. Koohikamali

Group 2:

Rahim Rashid

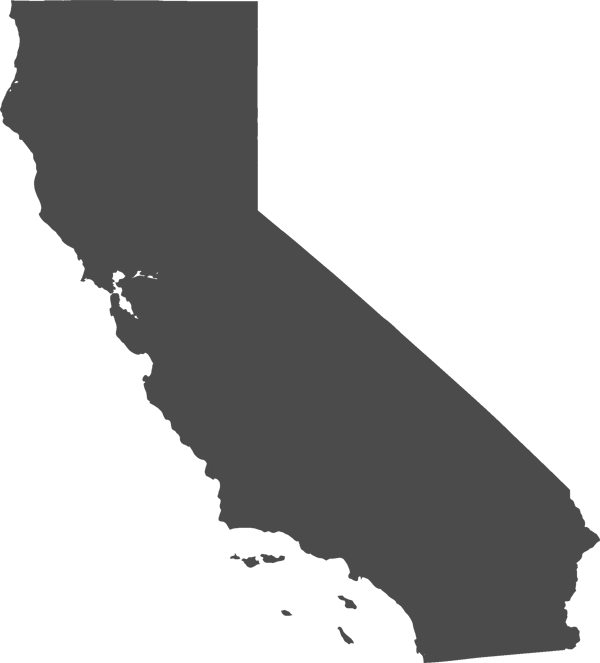
Dallas Venegas

William Vong

Jacob Yim

Contents

[Introduction 3](#_Toc89540686)

[Problem Statement 3](#_Toc89540687)

[Research Question 3](#_Toc89540688)

[Background Research 4](#_Toc89540689)

[Spatial Data Source 4](#_Toc89540690)

[Methodology Flowchart 5](#_Toc89540691)

[Spatial Analysis Method 6](#_Toc89540692)

[Results 6](#_Toc89540693)

[References 7](#_Toc89540694)

# Introduction

The following report includes the analysis of the collection and cleaning, exploring, and visualizing data from following data sets extracted by websites, csv files, and their conversions using geoprocessing techniques to display visuals and layers of geocoded data. Upon the completion of these tasks is where we perform our analysis techniques to derive information from our research. Our results will then be published as multiple webmaps as supporting evidence to our insights.

# Problem Statement

Developing foresight for major indicators of a disastrous event is complex in nature and often difficult to determine. Elevated blood lead levels are a serious threat to a community and may often go unnoticed should water districts go astray from protocols, as well as other unknown and unsuspecting threats which may assist in the elevations of blood lead levels. This report is to analyze and measure blood lead levels by county within California in the given year of 2019. Furthermore, we utilize our data from which we’ve capture and measure it against published data from our findings to determine possible sources which best addresses our statement.

# Research Question

Given the complexities of nature and the body of waters available to the state of California, our group will source, research, import, prepare, and clean our data to uncover the quality of data by county within California within the year of 2019 if possible. We will perform the same analysis on data pertaining to the blood lead levels in children and other possible sources of hazardous waste chemicals (or any substance), where the main objective is to possibly uncover a correlation between these variables. With our data, we intend to perform a spatial analysis from the data we have collected and will then geocode our findings to possibly find any visual correlations using spatial analysis techniques and such as clustering.

Should we find an increase in blood lead levels and decreased water quality, we may refer to the studies of the incident of Flint Michigan in 2014 and look for possible indicators from Flint Michigan and compare with the state of California. The goal is to explore data sets to find possible patterns and hidden structures within a large network in our data sets.

# Background Research

Similar research has been performed regarding the disaster of Flint Michigan in 2014 which led to a water crisis for the entire city. Residents of Flint Michigan began experiencing skin rashes, hair loss, and itchy skin prior to knowing that their water supply was contaminated. The source of the disaster stemmed from the city’s mismanagement of resources matched with corruption. According to our research, samples from Flint Michigan reached the “action level” of 15 parts per billion. It is most unfortunate for the residents of Flint, for nearly 9,000 children were supplied with lead-contaminated water for over 18 months.

Although the California has yet to experience such a disaster of mismanagement and corruption, we may use Flint Michigan as a baseline to our research and understand the precursors which lead to the disaster, and apply our findings at the city, county, or state level. This analysis will be determined upon our results.

# Spatial Data Source

Given the confidentiality of published data for water quality and blood lead levels for California, sources data sets are very scarce. In order to continue and perform a thorough analysis, it was required to extract tables from published credible websites by using python packaged libraries to create our own data frames and output our cleaned data into a csv file. Sources included the aggregation of blood lead levels for ages 21 and younger, as well as different tiers of blood lead level contamination ranges of less than 4.5%mcg, between 4.5% and 9.5% mcg, and over 9.5% mcg.

In addition, we were able to extract our resources to use point layers using XY coordinate data sets to plot our water districts within ArcGIS Pro. This step required the conversion of a data from a CSV file and undergo geoprocessing of the data to use a workable web map.

Preparing these two data sets for use in ArcGIS Pro was most essential to our analysis given the lack of current published data. Upon the completion of these tasks, we began to research additional variables by searching across ArcGIS’ sources to find additional layers which may bring further insights to our research.

# Methodology Flowchart

**Collecting and cleaning of Data**

**Better Perspective**

**Research on Background**

**Positioning**

**Explorative Analysis**

**Areas of Interests**

**Discover**

**Develop Insights**

* Credible
* Web Scraping
* Cleaning/Formatting
* Geoprocessing
* Understanding blood lead levels by CDC standards
* Research historical events in Flint, MI
* Exploring features which may elevate BLLs
* Utilizing tools in ArcGIS Pro
* Geoprocessing points of interests
* Add additional feature layers
* Examine features across BLLs
* Perform Analysis techniques of geoprocessing & clustering
* Develop a Dashboard
* Complete and publish webmaps

# Spatial Analysis Method

After we have explored our data and come to understand the definitions of our variables, we then prepare ourselves to perform our spatial analysis methods. Given that we have extracted our data from published websites, (where government websites have redacted results), we then layered the necessary features to our map and have decided that an analysis for clustering and grouping method would be most fitting. The purpose of this analysis is to understand possible organizations of unlabeled data into similar groups, known as clusters. A cluster itself is a collection of data which are similar and dissimilar to data items in other clusters. Data clustering is a method of unsupervised machine learning where data is separated into groups or clusters based on a similar measure, as well as determine if our data has natural groupings. We also attempted two other spatial analysis methods that were tasks and operations dashboard and geoprocessing. In our dashboard, we included two pie charts, one bar chart, one heat map, and our final layered map.

# Results

Map

Description automatically generatedMap

Description automatically generatedGiven the results of our findings, it would appear that the metropolitan cities and counties had the most counts of blood lead levels beyond 9.5mcg were Los Angeles and Sacramento. These counties have the largest populations (denoted in the red/pink gradient scale layer). Indicators that we may want to observe would be smaller populations with greater BLLs, such as counties to the east which experience the higher BLLs with smaller populations. Furthermore, we observe our centroids and clusters for pesticides and solid waste centroids. Our cluster analysis displays the different centers of groupings of solid waste and our clusters or pesticides. We can see that these layered elements include larger BLLs where clusters of pesticides and solid waste increases, however the results may also determine that larger populations and their exposure to BLLs may scale. Lastly, due to scaling and spatial relationships we must keep in mind of not stretching the map or else the connections between the layers and analysis will be incorrect.

# References

**Research Notes & Data Sources**

1. Gorr, W. L., & Kurland, K. S. (2020). *Gis tutorial for ArcGIS Pro 2.6*. Esri Press.
2. California Blood Lead Data 2019. (n.d.). Retrieved December 2, 2021, from https://www.cdph.ca.gov/Programs/CCDPHP/DEODC/CLPPB/Pages/BLLMapsTables2019.aspx.
3. *California County boundaries*. California State Geoportal. (n.d.). Retrieved December 2, 2021, from https://gis.data.ca.gov/datasets/CALFIRE-Forestry::california-county-boundaries/explore?location=37.246136%2C-119.002032%2C7.21&showTable=true.
4. California Blood Lead Data 2019. (n.d.). Retrieved December 3, 2021, from https://www.cdph.ca.gov/Programs/CCDPHP/DEODC/CLPPB/Pages/BLLMapsTables2019.aspx#.

**ArcGIS Published Content from members of the group:**

1. Dashboard:

<https://www.arcgis.com/apps/dashboards/77492787560d456fbebd98f1cc25f7a6>

1. Notebook:

<https://pomona.maps.arcgis.com/home/notebook/notebook.html?id=a404804e44174a4a9ded1b7b9a62b65f>

**WebMaps**:

1. Cluster Centroid Population & BLLs:

<https://arcg.is/18ueeu>

1. Spatial Aggregated Sum Heat Map:

<https://arcg.is/1WyiOT>

1. CDPH Blood Levels Plot:

<https://arcg.is/0uvDiu>