The Spatial-Environmental Dimensions of Cardiovascular Disease in California

---

Jack Daley
8 December 2023

Abstract: Cardiovascular Disease is a serious issue in California. While many immediate causes of this disease stem from internal health conditions in people, external factors have also been proven to increase CVD likelihood. Some of these stem from changing environmental conditions mainly driven by human activity. Examining data from the California Office of Environmental Health Hazard Assessment, I analyze which environmental factors may be affecting the risk of CVD amongst populations in California. I introduce a spatial dimension to the study; analyzing where certain variables may predict CVD better than others. In addition, I examine which variables are most important overall in predicting this risk, and how to optimize a model that best fits the data.

Keywords: Cardiovascular Disease (CVD), Random Forest Regression, Feature Ranking, Spatial Analysis

Introduction:

Cardiovascular disease (CVD) has proven to be a serious health concern globally. Risk factors for heart diseases vary from high or low blood pressure, high or low cholesterol, diabetes, excessive smoking, to trends of physical inactivity. The CDC reports that a striking 1 in 3 deaths in the United States are caused from cardiovascular disease. In California, the CDC reports that CVD is the leading cause of death in 2017 as well as the two years prior, surpassing the leading cause of cancer from 2014 to 2015. These striking statistics indicate that California is no exception to the devastation caused by this health problem (CDC, 2018).

Examining the environmental causes of vulnerable medical conditions leading to heart disease can be an effective strategy to understand CVD. Air pollution from primary and secondary source pollutants including Carbon Monoxide, Nitrous Oxides, and Ozone have all been shown to increase risk of CVD. Certain metals present in drinking water, including arsenic and lead, have also been linked to CVD. Finally, traffic has been proven to be associated with CVD, although it is not known whether the air pollutants or mental stress behind this factor contribute more (EPA, 2009). It is clear through this examination that CVD is caused by immediate medical conditions that can be exemplified by changing environmental conditions. The degree to which some of these environmental conditions heighten the risk of CVD appears somewhat unknown. This study aims to quantify how certain environmental factors affect CVD risk across space in California.

Methodology & Data:

It is apparent that many variables play a role in a population's risk of CVD. Using data collected amongst spatially defined populations can help to determine how much of a role these variables play. The California Office of Environmental Health Hazard Assessment has conducted

an ongoing project on the environmental and human health of California over the past decade. Known as the California Environmental Screening, or CalEnviroScreen, this study measures a number of variables over defined census tracts throughout the state of California. Each of these features is one of four themes: environmental indicators, environmental effects, social characteristics, or indices that combine a number of features. Considering that only one numerical value can be measured for each census tract, various geospatial techniques such as spatial autocorrelation are used to average values spread out over each tract (OEHHA, 2023). Updated versions of the CalEnviroScreen dataset are published every few years.

The CalEnviroScreen contains a spatial dimension to the dataset with geographic coordinates for the centroids of each census tract. These longitude and latitude coordinates allow for spatial analysis to be conducted on certain features. Visualization techniques can help determine which features are most apparent in what locations.

For my analysis, I chose to use the CalEnviroScreen dataset to examine risk of cardiovascular disease occurrences in relation to a number of factors. The ‘cardiovascular disease’ feature (my independent variable) was measured by taking the number of emergency department visits from the Office of Statewide Health Planning and Development (OSHPD). The other factors, chosen from the dataset, are as follows: PM 2.5 (particulate matter of size 2.5 nm), Diesel PM, Drinking Water, Lead, Pesticides, Toxic Releases, Traffic, and Hazardous Waste Sites (my dependent variables). Each of these features were selected because they correspond either directly or indirectly with known causes of cardiovascular disease. In addition, I chose these particular variables to try to differentiate between urban (such as ‘traffic’ and ‘drinking water’) and rural (such as ‘pesticides’) environmental concerns.

I began with structuring the data and visualizing the data. Using pandas dataframes and matplotlib, I was able to spatially visualize and compare cardiovascular disease with a few other variables. This analysis was similar to typical geospatial methods of analysis in ArcGIS Pro.

Upon visualizing the data, I engineered the features by checking for non existing values, categorical variables, or other potential problems for the model. I quickly fixed these issues. I then split the data into training and testing sets based on a test size of 0.2, or 20% of the data. I used this split because the CalEnviroScreen dataset contains sufficient data for training and testing.

To estimate CVD occurrences, I decided to use a Random Forest Regressor model implemented with all of the selected independent variables. Upon shifting the model parameters a significant number of times, the ensemble forest regressor obtained favorable results with around 800 estimators, a maximum depth of 20 nodes, and 8 maximum features per decision tree. This forest obtained an r squared error metric of approximately 0.54, meaning that it was able to explain a little over half of the randomness in the data. Computational efficiency declined significantly when creating any forests larger and/or deeper than this, so I decided that this was the optimal model for my study.

Upon construction of my model, I began analyzing the model's performance and accuracy. For my study, it was important to not only analyze the model efficiency on an overall basis by measuring the regressor score, but also to analyze individual errors spatially. I chose to take four geographic subsets of the dataset, two in the urban parts of California and two in the rural parts of California. For the urban datasets, I chose to analyze the Los Angeles Metropolitan Area and the San Francisco/Sacramento Metropolitan Area. I chose these areas for several reasons, one being that they are the two largest and densest metropolitan areas in the state, with

favorable environmental conditions for CVD. Another is that many variables tend to have much significance in these areas. After selecting these two urban areas, I selected two rural areas of the state. One is the San Joaquin Valley approximately from Fresno to Bakersfield, and the other is the Monterey Bay and surrounding farmlands. It is important to note that the geographic subsets that I construct are rectangular, so the primarily 'urban' regions contain rural areas around the edges, and the primarily 'rural' regions contain small urban areas within them.

After creating the model, I continued to analyze the data using a feature ranking methodology. This numerical approach to information gain uses mathematical processes such as entropy and GINI impurity to calculate the bits of information gained from individual features in relation to another. This kind of analysis was necessary in my study to be able to analyze which of my input features explained more of the randomness in the dataset than others. I intended that this analysis would help to explain which potential causes of CVD are most important to study.

Results:

The results of this study range from visual analyses to graph and histogram analyses. Fig 1 shows a distribution of risks of CVD across the state of California. Clear distinctions between areas of high CVD risk and areas of low CVD risk are shown.

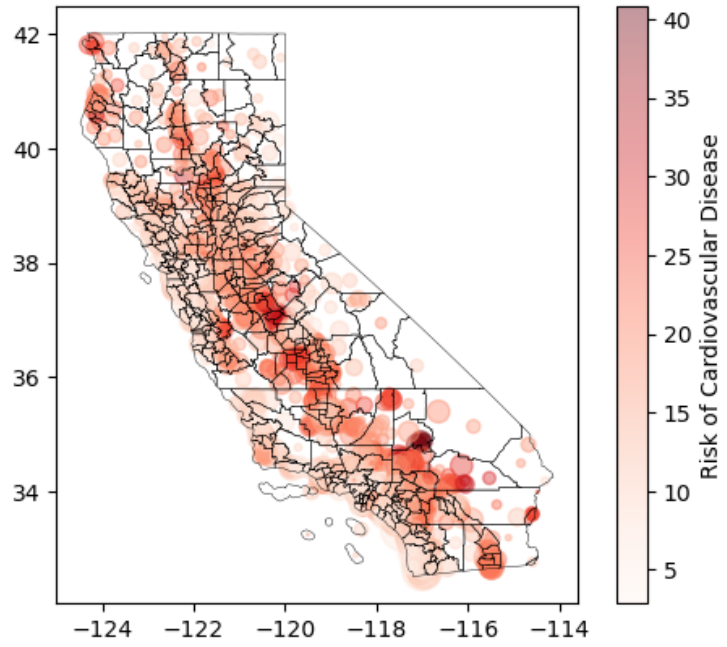

Fig. 1 Spatial Analysis of CVD occurrences across California.

Two maps of independent variables, lead risk and the logarithm of Pesticide Releases, were produced to visualize. The lead risk variable shows particularly high values in urban areas of California such as Los Angeles and the San Francisco Peninsula. The pesticides distribution shows particularly high values in rural/agricultural parts of the state such as Ventura, the San Joaquin Valley, and the Monterey Bay Area. These differences hint at how the model could estimate CVD differently in different locations.

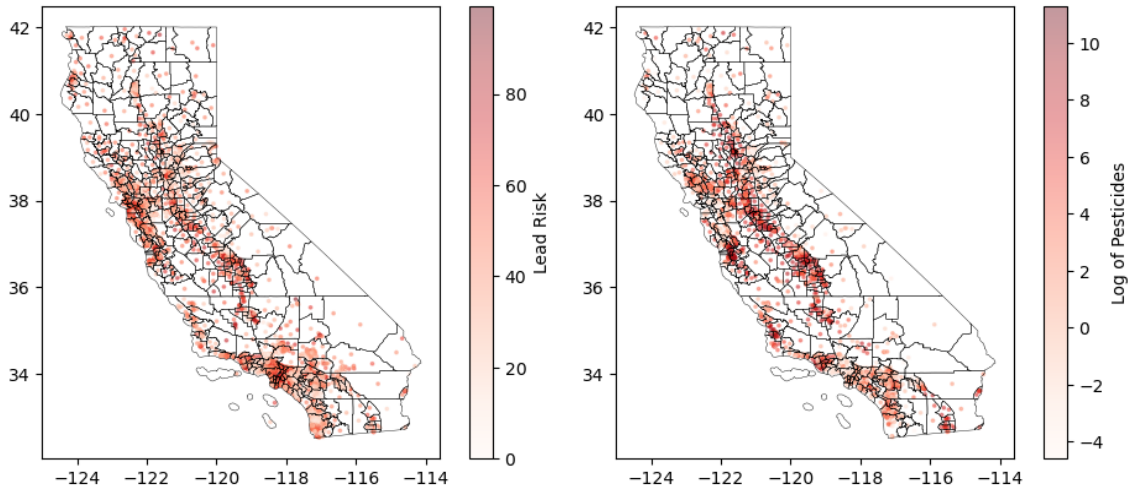

Fig 2. Analysis of Lead Risk and Logarithm of Pesticide releases across California.

Also using matplotlib, a figure was created showing four maps to visualize CVD risk across the four selected regions of study around California (see fig 3). These four maps reveal interesting results about the dependent variable, CVD. They show that in urban areas, such as Los Angeles and San Francisco, CVD risk is not evenly distributed. In Los Angeles CVD risk is high in Central and Southern LA, whereas it is low in surrounding areas such as West LA and Orange County. In San Francisco, CVD risk is high in areas such as Richmond and Fremont, whereas it is low in areas like the SF Peninsula and Marin County. In the rural maps, CVD risk appears to be especially high. The east part of Plot 2 shows the Sacramento Valley, where CVD risk is very high. Plot 3 shows the Central San Joaquin Valley, where remote rural areas appear to have very high CVD risk. In addition, plot 4 shows agricultural areas around Monterey Bay, where CVD risk appears variable but very high in some eastern locations.

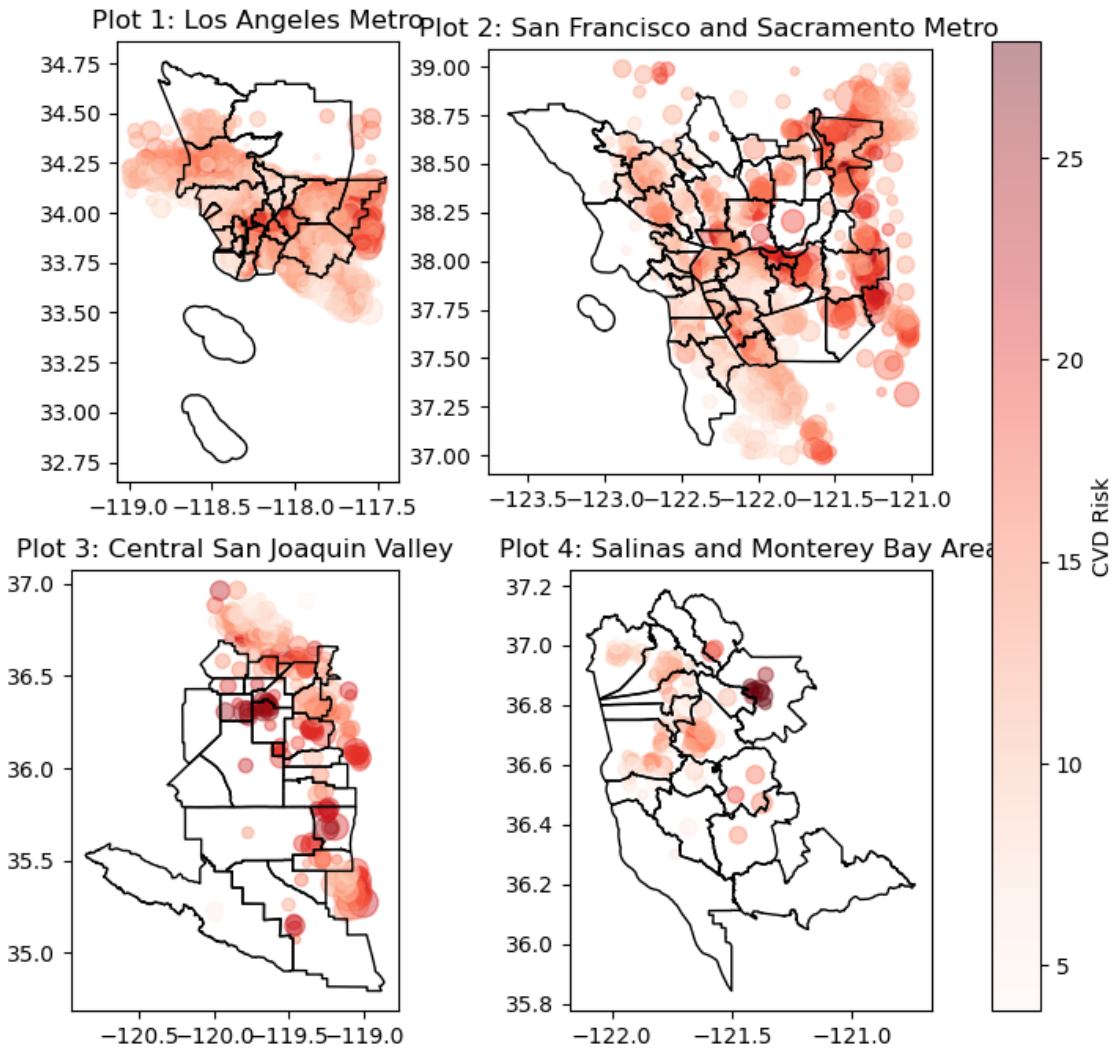

Fig 3. Spatial Distribution of CVD Risk

Upon running the random forest regressor model a number of times, an error assessment was conducted. All statewide spatial distributions of errors were comparable to figure 4. An example of study area specific errors is shown in figure 5. This figure revealed varying results, however some trends can be made out. In urban areas where CVD Risk is low (such as San Francisco and West LA), errors consistently across runs of the model tend to be low. In urban areas where CVD risk is high (such as Sacramento, Fresno, and Bakersfield), errors tend to be

high. In rural areas, such as to the East of the Bay Area, the San Joaquin Valley, and the Monterey Bay Area, errors are variable but have some extremely high values.

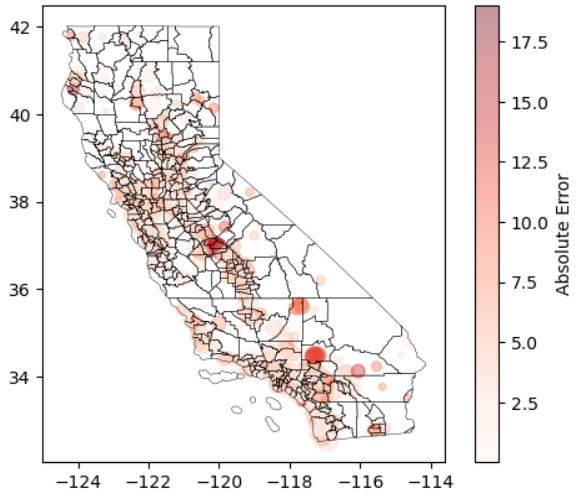

Fig 4. Spatial distribution of errors across California.

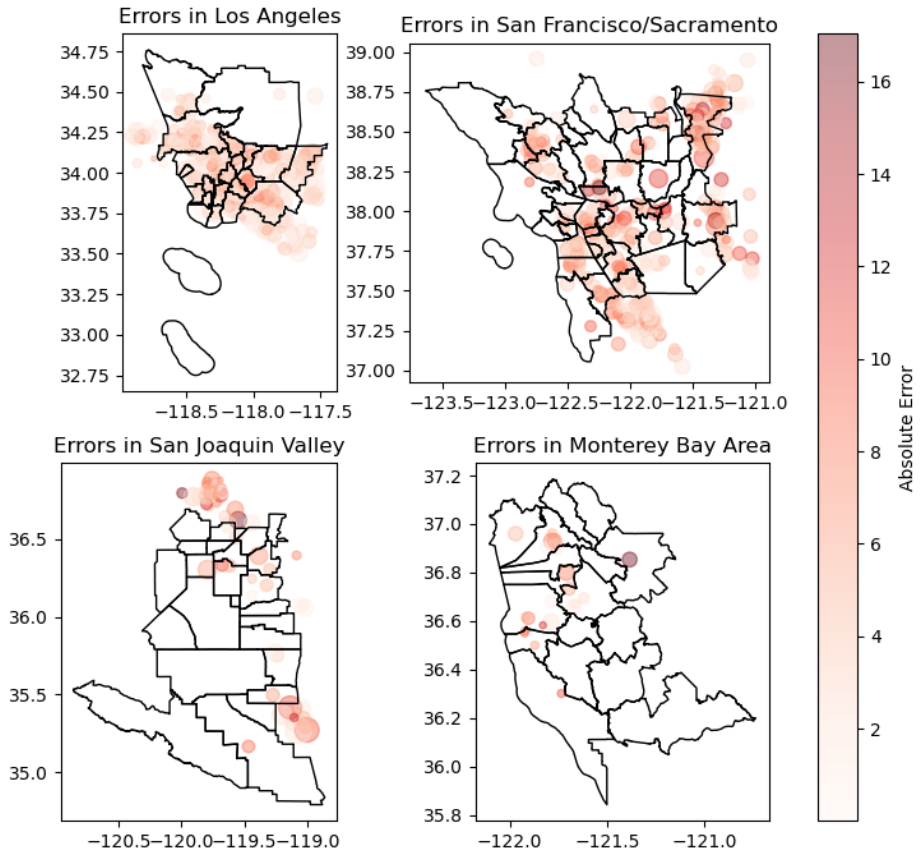

Fig 5. An example of the spatial distribution of absolute errors produced from a Random Forest Regressor Model

Region	Average Error
Los Angeles	2.43
SF/Sacramento	2.339
San Joaquin Valley	2.669
Monterey Bay Area	2.175

Fig 6. Average errors amongst regions, relatively consistent across model runs

After all of this modeling, the random forest regressor was used to rank features. A table is shown containing each feature and its feature rank, explaining which of the features helped to predict CVD the most and which helped to predict it the least (see fig 7). Running the model a number of times resulted in consistently similar feature importances. A bar chart was created to more easily visualize the distribution of these feature rankings (see fig 8).

Measure	Feature Importance
Toxic Releases	0.2056
PM2.5	0.2044
Drinking Water	0.1683
Lead	0.1645
Traffic	0.0811
Hazardous Waste Sites	0.0617
Pesticides	0.0433

Fig 7. Feature rankings of each of the chosen features for the random forest model, relatively consistent across model runs

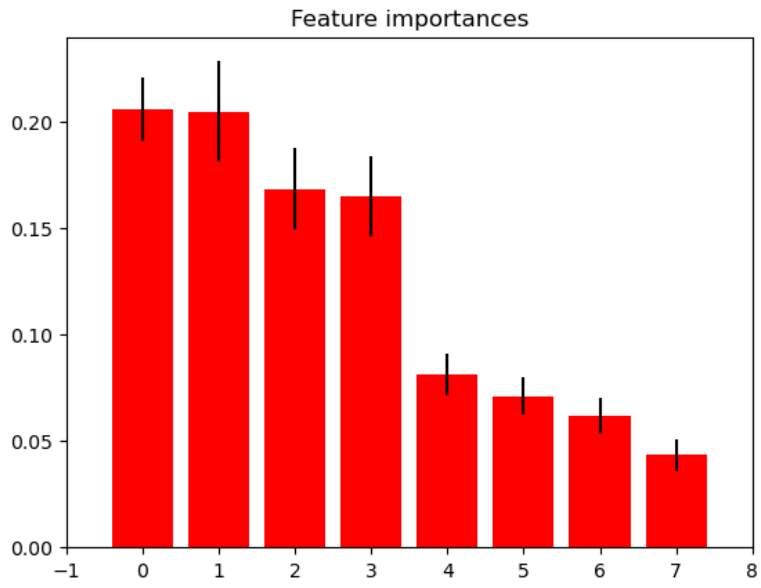

Fig 8. Bar Chart of feature rankings. Numbered bars correspond to previous row numbers.

Discussion:

This study provides intriguing analysis into how we can measure the risk of CVD across California and perhaps beyond. When visualizing the distribution of CVD occurrences in the state, the distribution at first appears quite random. However with a dataset such as the CalEnviroScreen, much of the randomness can be explained by other measurements.

Beginning with the selection of features, the CalEnviroScreen dataset provides measurements of causes and effects of environmental harms in attempts to identify vulnerable communities in the state of California. It prefaces that harmful environmental effects are not spread evenly across the state, thus it chooses variables that clearly show geographic skew in certain areas. This selection of variables to include in the dataset may have undesired effects on my model, where skewed data in disproportionately affected communities change the model predictions in unintended ways. In general, this uncertainty suggests that choosing the correct variables is vital to constructing an optimal model to predict CVD risk.

One important analysis that my model conducted was where in California the selected variables could predict CVD the best and where the model lacked the most. As analyzed in the results, the spatial distributions of errors can be linked to characteristics of the dependent and independent variables. The correlation between rural versus urban CVD risk and error leads me to believe that I included too many features that are primarily present in urban areas (such as traffic, Diesel PM, PM2.5, and Lead risk) and not nearly as present in rural areas. Given that errors tend to be a bit higher in remote rural areas, I believe there may be more factors leading to CVD risk in rural areas that the model does not account for. Perhaps if I had researched more indicators of CVD primarily present in rural areas such as mineral dust air particle concentrations the model would have performed better in these areas.

When examining the features rankings that were derived from the model, interesting conclusions can be drawn. Toxic releases and PM2.5 were determined to be the variables most correlated with CVD. These variables are both associated with atmospheric conditions, confirming that CVD risk is primarily determined through respiratory factors. Drinking water and lead were the next two ranked features which were interesting results to me considering these measures involve another biological system. However this aligns with the EPA assessment of risk, mentioned in the introduction. The fact that they are still correlated with CVD risk indicates that they still could be used for predicting certain health conditions. I was also surprised by the ranks of the ‘traffic,’ ‘hazardous waste,’ and ‘Diesel PM’ variables. I predicted these variables to explain almost as much randomness in CVD risk as PM2.5 considering they are also associated with atmospheric conditions. I predict that this again may have to do with the urban/rural divide: the three low ranking features are primarily urban based and thus create high

error in rural areas. Figure 6 further confirms this, as the San Joaquin Valley, one of the largest rural areas in the state, consistently ranks highest in error out of all the study regions.

Despite all of these shortcomings of the random forest regressor model on this dataset, the model still performs well, especially in certain areas, and could be a reliable predictor of CVD risk. With more public health research and more defined study areas, this process could be refined and improved significantly to inform future socio-economic decision making.

Works Cited:

"CalEnviroScreen 4.0." *Oehha.ca.Gov*, OEHHA, 1 May 2023, oehha.ca.gov/calenviroscreen/report/calenviroscreen-40.

"Stats of the State of California." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 13 Apr. 2018, www.cdc.gov/nchs/pressroom/states/california/california.htm.

"United States Environmental Protection Agency." Aug. 2009.

Data Sources:

California Office of Environmental Health Hazard Assessment, TIGERLINE Shapefile

Software:

Jupyter Notebook, Python, Excel