# Estimate Population Health at Census Tract Level Based on Crime Data

**Yiqiao Li**
University of Southern California
yiqiaoli@usc.edu

**Abstract**

Does your zip code affect how healthy you are? Resident health has always been city planners' and local health department's top priority. However, limited health data analysis is available, partly due to the difficulty in obtaining and analyzing both a large scale data for cities and small areas data within cities. We address this problem by exploring datasets from 500 Cities project data from CDC and crime data from LAPD and find correlation between these datasets to better understand geographic health distribution. We predict the health outcome by learning supervised regression models. Furthermore, the prediction accuracy can be significantly improved by performing various preprocessing. Although the accuracy is not optimal, we believe our finds on interplay of crime rates in census tract and health outcome in that region provide a new perspective on predicting resident health.

## Introduction

Population health study has been conducted for decades. Researchers, city planners and local health departments all want to better understand the burden and geographic distribution of health-related variables in their jurisdictions. Commonly, neighborhood food environment, air pollution, water pollution [1] etc. are all considered as key factors that affect resident health. The strongest predictors of health in multivariate and multilevel models were income, trust in politicians and governments, and trust in other members of the community [2]. Community-level study also indicates that neighborhood influences people's health outcome [3].

As a result, a healthy and trustful community is very important to the well-being of residents. I believe safety is one of the key roles of a healthy community. Thanks to pervasive open data nowadays, we can leverage machine learning techniques to discover potential patterns between different factors. For this reason, I conduct this study on finding correlations between unsafe behavior, especially violent behavior, namely crime, and health outcomes.

Copyright © 2018, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

## Methods

### Datasets

I use three datasets including LAPD crime dataset[4] which reflects incidents of crime in the City of Los Angeles dating back to 2010, Census 2010 - Census Tract dataset[5] from Census Bureau and 500 Cities Data[6] which provides city and census tract-level small area estimates for chronic disease risk factors, health outcomes, and clinical preventive service use for the largest 500 cities in the United States.

| | |
|---|---|
| Crime data | 2010 - present |
| | 194889 instances |
| | 26 attributes |
| Census tract data | 2014 |
| | 1005 instances |
| | 24 features |
| | Physical health not good for >=14 days |
| Census boundaries data | GeoJSON |

*Figure 1. Three datasets*

### Data merging

First of all, I use the Census Tract dataset to define the boundaries of all small areas of the Los Angeles city. For each area, I find its health outcome value from 500 Cities health dataset by seeing if the geo points could fall into a small area. As a result, I have geometry shapes containing its own health value for each small area. Furthermore, I map all the crime data to the spatial dataset I generate from the second step.

In order to ease the merging of these three datasets, I perform a three-way spatial join by using geolocation features as the keys. I use a C-programming package - Spatialindex to speed up the three-way join, since spatial join is extremely time and space consuming.

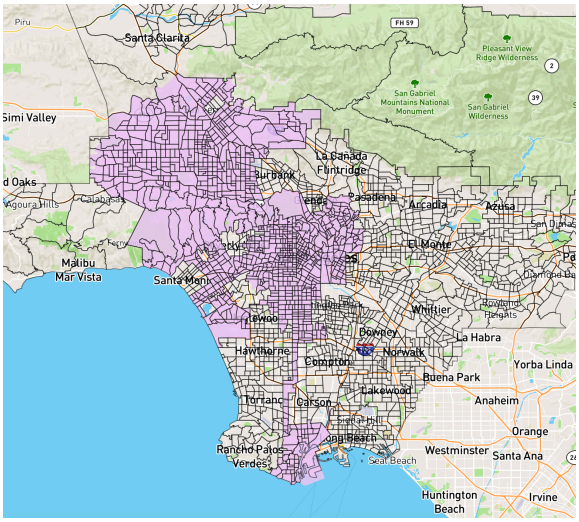


*Figure 2. Census tract mapping*

## Tools

Spatialindex, Python 3, Scikit-Learn, Jupyter notebook, Pandas, Numpy Scipy

## Feature Selection

After joining the datasets, we have a huge dataset containing 194737 instances with tens of thousands missing values. If we just drop all the missing values, we end up having only 59405 instances and 9 features and the prediction accuracy suffers a lot due to the lack of features since some of the 9 features are highly correlated already, such as Area name and Area code, etc.

Since the dataset has a lot of missing categorical value, I use Imputation transformer to completing missing values and use One-hot encoding method to create one binary attribute per category. After one-hot encoding I get a matrix with thousands of columns, and the matrix is full of zeros except for one 1 per row. Because our dataset is very large and using up tons of memory mostly to store zeros would be very wasteful, I instead use a sparse matrix only stores the location of the nonzero elements for each category.

## Feature Scaling

After feature selection, I have a large enough dataset with 16 features. I find that the machine learning algorithm doesn't perform well when input numerical attributes have very different scales. However, this is the case of my dataset: Victim Descent variable is around 141800000 while the Area Id is ranging from 0 to less than 100.

I use standardization method to get attributes to have the same scale since the method would not be much affected by outliers.

## Further preprocessing

I notice that some attributes have a tail-heavy distribution such as Age attribute. Most victims' age is below 40. There is a long tail when age ranges from 40 to 90. I transform the series with the natural logarithm to gain slightly better prediction accuracy.

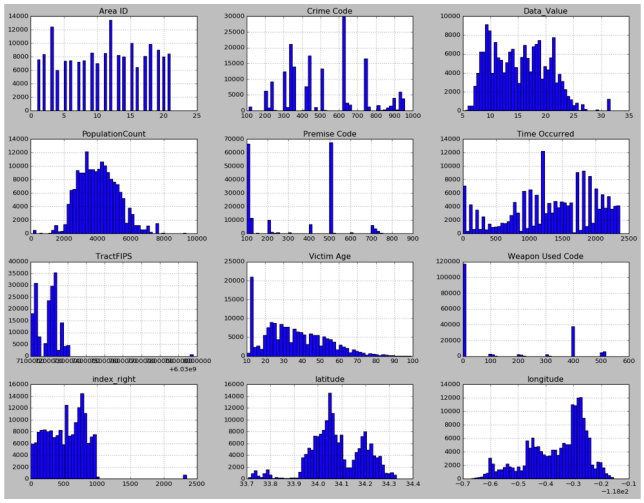


*Figure 3. Features distribution*

## Tools

Python 3, Scikit-Learn, Jupyter notebook, Pandas, Numpy Scipy

## Data Visualization

I use Mapbox and Matplotlib to visualize the result. Red points represent the crime locations while the green spots represent the health status of certain Census Tract.

## Tools

Matplotlib, Mapbox

# Experiment

## Training and result

I use Linear regression, Support vector regression, Random Forest regression as train models for both 9 featured and 16 featured heavy-preprocessed dataset respectively. Random Forests regression outperform others.

| ML Model | 9 features | 16 features |
|---|---|---|
| Linear regression | 42.7% | 59.1% |
| Support vector regression | 41.7% | 57.2% |
| **Random Forests regression** | **47.3%** | **68.4%** |

*Figure 4. Results*

I render the interactive map by plotting the crime incidents **related to assault** and health value of each census tract. The LA downtown and southern LA are the high violent crime rate area. The west LA and northwest LA are relatively safe.

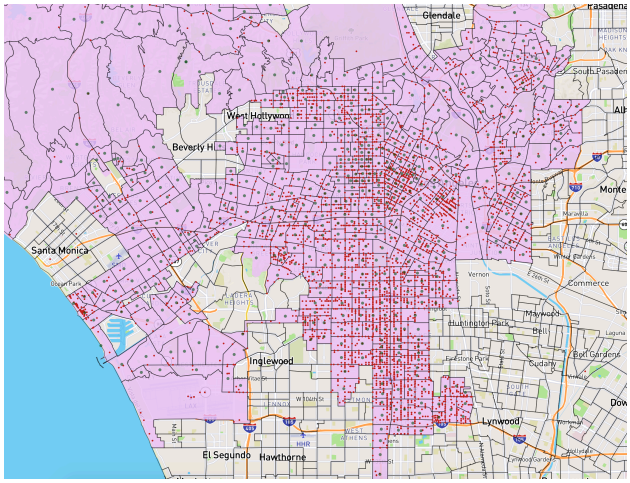


*Figure 5. Crime data*

We can observe that these maps are overlapped in terms of crime rate intense areas and unhealthy areas. But there is bias and inaccuracy in some areas. For example, Venice beach which is unsafe according to the crime map tends to be quite healthy.

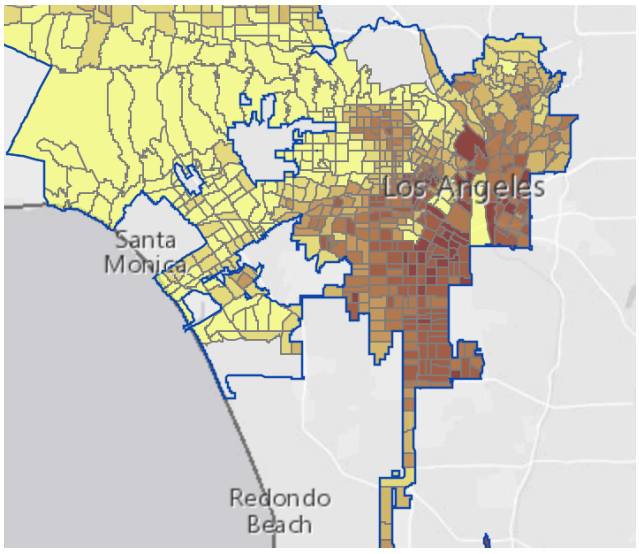


*Figure 6. Health values*

## Conclusion

Data preprocessing plays an important role for this dataset which has some chaos in it. Random forest tree outperforms other models since the model is unlikely linear. Finally, the result to some extent provides a new perspective on predicting population health using crime data.

In future, further feature engineering could be done to improve the accuracy. I think we can focus on studying how certain feature influences the result most. Also, more visualization can be done to further demonstrate the data geographical characteristic.

## References

[1] S Macintyre, A Ellaway, S Cummins - Social science & medicine, 2002. "Place effects on health: how can we conceptualize, operationalize and measure them?

[2] Gerry Veenstra. 2005. "Location, location, location: contextual and compositional health effects of social capital in British Columbia, Canada"

[3] Roy J. Shephard, 2008. "Is Active Commuting the Answer to Population Health?"

[4] https://data.lacity.org/, LA Open Data Portal

[5]https://www.cdc.gov/500cities/methodology.htm, methodology of 500 city data.

[6]https://www.census.gov/geo/maps-data/data/tiger-line.html, 2015 census tract shape file.