CS 591 L1 Report
Comparing Regions of Crime with Regions of Environmental Investments in Boston

Jonathan Silverman, Rohit Philip Mathew

Abstract

Examining crime levels are a vital part of understanding the extent of development of a neighborhood. Economists reckon that there exists a link between environmental investment and economic growth. In this period of heavy debate among politicians, lawmakers, and even citizens concerning the effects of pollution on the planet, we have decided to try and examine the extent to which neighborhoods in Boston are trying to reduce the size of their carbon footprint, and if there exists an observable link between crime levels with the number of environmental assets in that same area

Introduction

A common metric for measuring the overall quality of a city is, not surprisingly, the crime rate. Crime is a factor that statisticians and sociologists have linked to everything from overall police volume and effectiveness to income and school quality. By analysing these correlations, one can better understand the contributing factors behind crime and hopefully implement meaningful policy to mitigate its effects.

For our project, we decided to analyse the relationship between Boston's environmental infrastructure and its overall crime rate. However, recognizing that there are historically "bad" areas for crime, we realized that creating a one- size- fits- all model for the whole city would not yield entirely meaningful results. Thus we came to the conclusion that *k*- means clustering would be a good way to analyse this information.

Methodology

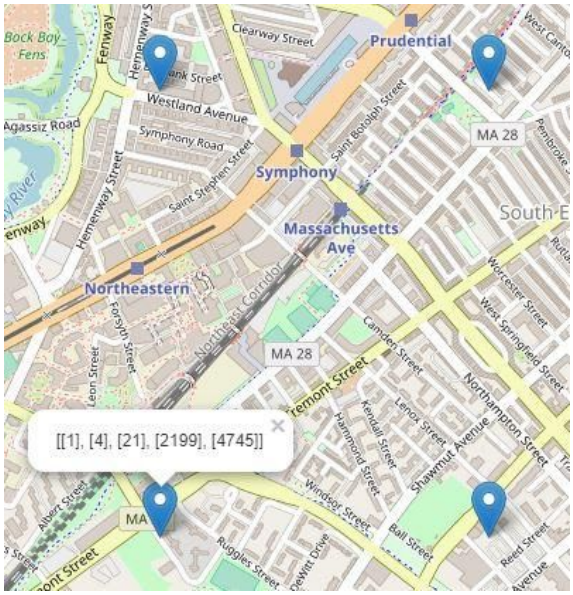
Data Collection

We collected data from five data sources. For analysing crime, we used the Crime Incident Report from 2015 onwards, which contained useful information about the location and type of crime. We also used datasets for the locations of environmentally- friendly infrastructure, what we will refer to as "Green Assets." These include open spaces, trees, charging stations, and Hubway stations. We used MongoDB to hold and manipulate the data.

Data Transformations

Data transformations for this project revolved around representing the collated data in a uniform manner that would make it easier to use. This involved cleaning some of the values, projecting certain features, and combining the resulting values. The Open Spaces in Boston dataset gave us the boundary coordinates of all open spaces (gardens, parks, playgrounds, etc.). Since we wanted to work exclusively with points, we represented these boundary values with just one pair of coordinates: the centroid defined by this boundary. At the end of this step, all our Green Assets and Crime locations were represented by a pair of values: the latitude and the longitude.

The next part of our transformation was an important step. Rather than working with these derived points directly, we decided to take inspiration from a Point Access Method called the Grid File, an advanced indexing technique. We divided the map of Boston into grids (the scale can be user-defined), and used the coordinates of each cell as keys, with a feature vector representation as its value. This feature vector was of length 5, representing each of the Green Assets and crime locations. The value of each dimension gave us the counts of the Green Assets/crimes contained in that cell. This method of partitioning greatly reduced the computation time for the Analysis stages, and also paves the way for future implementations when try to vary the granularity of viewing Boston from above. Consider the following subset of the entire grid that gives us the counts of Green Assets and crime incidents of the cell defined by the lower left coordinates:



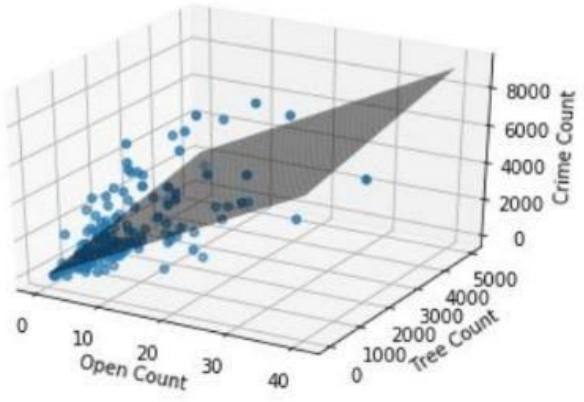
Please note that this differs from the actual Grid File implementation in the sense that it isn't dynamic, and once the scale is assigned, the size of the cells don't change.

Analyses and Observations

Method 1: Multiple Regression

After assigning each point coordinate from the database to their respective cell, we decided to try to find which Green Asset best correlated to the Crime Rate. Multiple Regression is a technique used to estimate the value of a dependent variable from a set of independent variables. In order to estimate the value of the predictors for each value, we used the Ordinary Least Squares (OLS) method.

After running this regression with Green Assets as the independent variables and Crime as the dependent, we observed each variable's confidence intervals to find which ones were significant. Through this, we were able to recognize that the confidence intervals for Hubway and charging station locations were far too wide. Consequently, we threw them out and ran the regression again with just the trees and open spaces counts. Here is a representation of the results:



Conclusion 1:

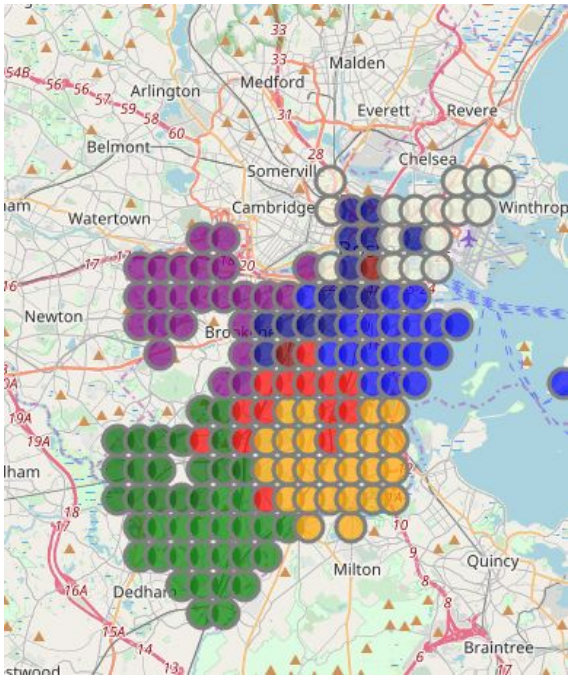
This model shows a positive correlation between crime, and open space as well as tree counts. It has an R- squared value of 0.675, meaning that the model fit the data decently well, in general. But moving forward, we wanted to include more of the geographical aspect of the data.

Method 2: *K*- Means Clustering

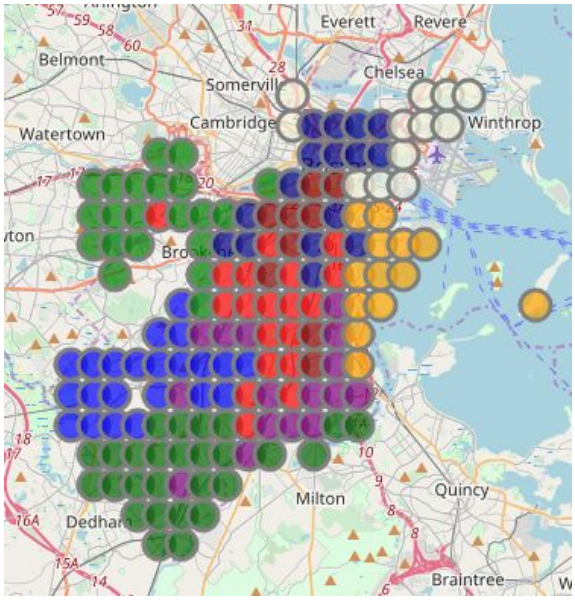
In order to incorporate the geographical aspect of our project, we figured the best way to do this would be to cluster our data points using *k*- means clustering, then compare the clusters for each Green Asset to the cluster for crime count. To set up the clusters, we first had to normalize all the count values and latitude and longitude parameters. Next we needed to find a *k* value that would work reasonably well with each set of values. We decided to use the Silhouette Score as the metric to help us figure this out. Silhouette score is a method for finding the "strength" of a cluster. The formula for a silhouette coefficient is: $S = \frac{b-a}{max\,(a,\,b)}$,
where a is the mean distance between a point and all the other points in the same cluster, b is the mean distance between a data point and the nearest cluster.

With this method, we were able to observe that for all the datasets, a local maximum existed with *k* to be around 10. This is the value we went ahead with. We also ignored the cells that had absolutely no data points contained within it, i.e.: the feature vector was made up of all zeros.Using

count, latitude, and longitude as our feature vector, we performed *k*- means for each item. Here is a representation of the clustering result for Open Spaces by count:



And here is a representation of the clustering result for Crime locations by count:



W used an Adjust RAND Index Score to compare the similarity of each pair of cluster results and found that the open spaces count cluster was most similar to the crime count cluster.

Conclusion 2:

The findings of the RAND Index score tell us that Open Spaces in Boston are closely linked to (or cause) Crime. We believe that this can be attributed to the simple fact that there are more open spaces in "bad" areas as compared to highly developed areas such as downtown Boston.

Future Work

In order to improve the accuracy and meaning fullness of our results, we could do one of various things. We could increase the granularity of the matrix beyond the 31 x 17 matrix (scale of 0.01) that we worked with, and repeat the experiments to try and increase accuracy. We also realize that since all our data came from different sources, the range that they cover varies greatly. For example, some of the datasets made no reference to Cambridge, while the rest encompassed data that went as far north as Everett. This will have most likely skewed our results, and we could work around this by setting stronger constraints during the pre- processing and transformation stages. Some of our data points included the Boston Harbour Islands, which are well beyond the mainland. Our grid included these values, and so, a large number of empty cells that contained nothing but the North Atlantic Ocean. Again, this would have skewed our results, and we could combat this by adjusting the boundaries of our matrix so that it is no longer rectangular, but trapezoidal; tapering in such a way that it minimizes the area covering the ocean as much as possible. Furthermore, we could perform more regression testing on each of the clusters to see if the global correlation is similar to the more local model. Similarly, to improve the accuracy of our regression testing, we could find a way to normalize the data based on population, foot traffic, or a similar metric.