BU CS591 Project Summary Report

A Characterization of Neighborhood Wealth and Optimization of Resource Distribution in Boston

Cody Karjadi and John Gonsalves

Introduction:

As a city grows larger, the task of distributing resources becomes increasingly complex. We propose that the process of resource distribution to improve infrastructure[1] can be optimized. The process can be broken down into weighted points of data and then this data can be fed into the appropriate algorithm. However, a characterization of the current resource status of a given area must be created before even beginning the process of analyzing resource distribution.

In this project we split up Boston by zip code, quantified current resource status as average property value, and the resource we examined was the potential building of new hospitals.

First, we attempted to quantify how accurate property value can be as a metric for current resource status by analyzing correlations between other factors in a given area and the average property value. To accomplish this, we retrieved many datasets https://data.cityofboston.gov/.

The datasets we retrieved included: community gardens,[2] crime incident reports,[3] food establishment licenses,[4] food pantries[5], hospital locations,[6] property assessments,[7] and Waze alerts.[8]

We then generated several datasets: **average property values** (total property value / number of properties in a given zip code; generated from **property assessments**), and **new hospital locations** (generated by algorithm; used **hospital locations, crime incident reports, and Waze Jam Alerts**).

The following datasets followed the paradigm of finding the correlation between (number_in_zip_code, average_property_value_zip_code) – for example, one data point could be (100,100,000) and could mean there have been 100 crimes in a given zip code, and this zip code average property value is 100,000. The datasets following this paradigm were: **community garden statistics, fast food establishment statistics, food pantries statistics, crime incident**

---

[1] **Infrastructure:** "The physical components of interrelated systems providing commodities and services essential to enable, sustain, or enhance societal living conditions."
[2] https://data.cityofboston.gov/Health/Community-Gardens/cr3i-jj7v
[3] https://data.cityofboston.gov/Public-Safety/Crime-Incident-Reports-August-2015-To-Date-Source-/fqn4-4qap
[4] https://data.cityofboston.gov/Permitting/Active-Food-Establishment-Licenses/gb6y-34cq
[5] https://data.cityofboston.gov/Health/Food-Pantries/vjvb-2kg6
[6] https://data.cityofboston.gov/Public-Health/Hospital-Locations/46f7-2snz
[7] https://data.cityofboston.gov/Permitting/Property-Assessment-2016/i7w8-ure5
[8] https://data.cityofboston.gov/Transportation/Waze-Alert-Data/h8aq-6mw7

statistics, and hospital location statistics. All of these datasets used the respective original dataset from data.cityofboston.gov, and the average property value data we generated.

Methodology:

Our statistical calculations of correlation values and p-values followed the general standard of calculation. Correlation values were calculated as COV(x, y) / STD(x)*STD(y), and p-values were calculated as the percentage of permutated datasets that had correlation values greater than or equal to the original correlation value (e.g., p value of 0.05 interpreted as 5% of randomly permutations of the dataset that had higher correlation values). The scalability of the correlation calculations may be difficult as the datasets increase in size, as that increases the time it takes to create some N permutations of the dataset (currently N = 4000 for our calculations). The usefulness of this statistic may not solve any problems directly, but provides a decent first step. If a correlation is found to be significant (p value <0.05 or so), we can only say that we observed some correlation and the chance that a random permutation has a higher correlation is very low. This doesn't necessarily translate into any solid usable information, as other factors can be influencing the data that aren't taken into account.

Our algorithm for new hospital locations begins by clustering Waze jams and crime incidents to their respective closest hospital. The crime incident reports had only police districts listed, which we then converted to zip codes[9]. This temporary dataset for a given hospital looks like (latitude, longitude, Waze jams+ crime incidents). Therefore, every Waze jam and crime incident has been grouped with the closest hospital.

This temporary dataset was then fed into a 3D k-means algorithm. K-means will start with some k initial means, and every observation will be clustered to the closest mean. Then, every observation is used to recalculate the respective mean it was clustered to (sum of all x / # observation of x, sum of all y…, sum of all z…). This is repeated until new means are no longer able to be calculated (the means from the last iteration = the means from newest iteration; this means that no more means can be calculated).

The k means that are calculated are interpreted as the most optimal locations for building k new hospitals, based on the metric of location, crime incidents, and traffic jams. We propose that the k means we calculate will optimally distribute the weight of location, crime, and traffic jams. The latitude / longitude were extracted from the final means, and then displayed on our visualizations.

One potential issue for scalability for this algorithm may be the clustering. For any given point, its distance from every single existing hospital is calculated. If the number of points or number of hospitals becomes really high, the clustering may add a lot to the calculation time.

The main issue with this algorithm is that the outputted latitude/longitude coordinates may not necessarily be feasible. For instance, that location could be another building or in the middle of a road. However, the main goal we wanted to accomplish was creating a paradigm to optimize resource distribution. Our algorithms show that this is possible by using the simple idea of

[9] http://bpdnews.com/districts/

clustering and then executing k-means calculations. With more refined datasets and "realistic" boundaries, this sort of paradigm could be very useful.

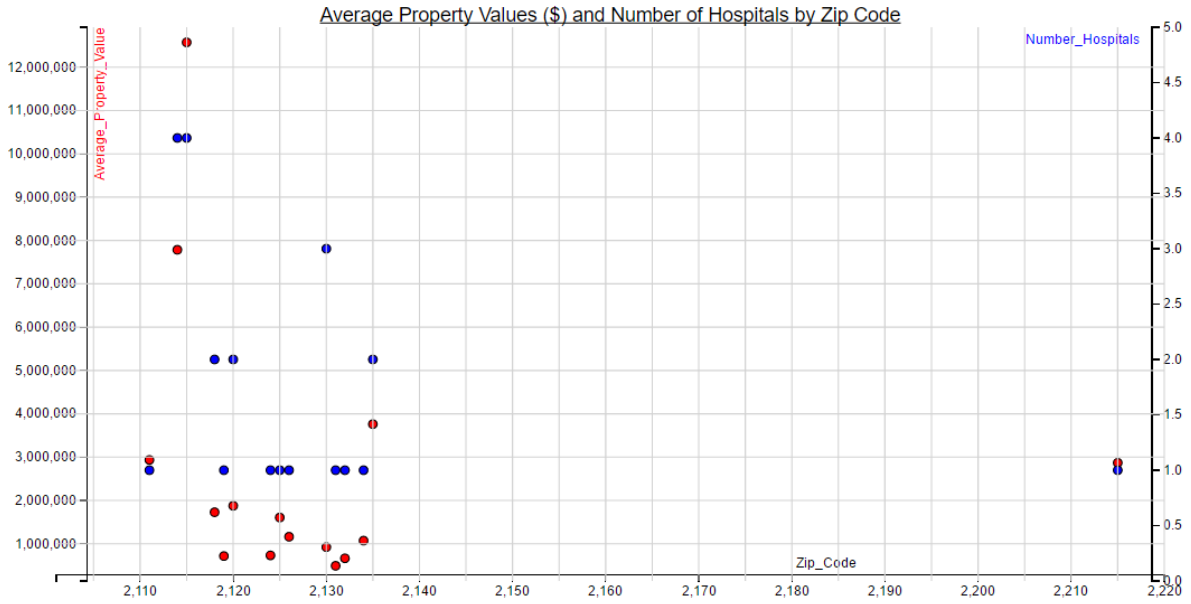
Results:

The following table lists our correlation and p-values we calculated for each dataset, comparing each dataset to **average property values.**

Dataset	**Correlation**	**p-value**
Community Gardens	-0.28	0.943
Crime Incidents	-0.172	0.674
Fast Food Establishments	**0.358**	**0.043**
Food Pantries	-0.355	0.958
Hospitals	**0.793**	**0.0027**

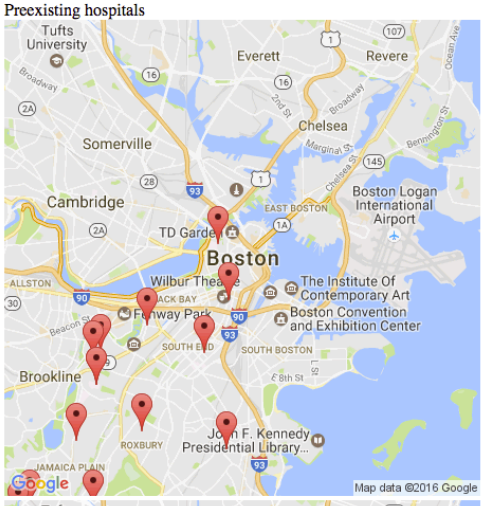
The only correlation values found to be significant were hospital locations and fast food establishments; meaning that we observed high property values in the same zip codes where there were high numbers of hospitals and fast food establishments, respectively.

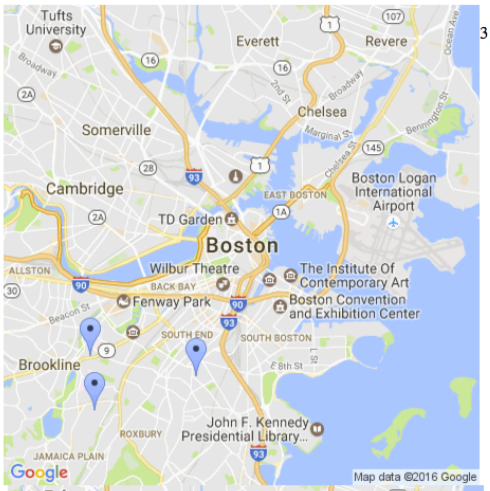
Visualizations:



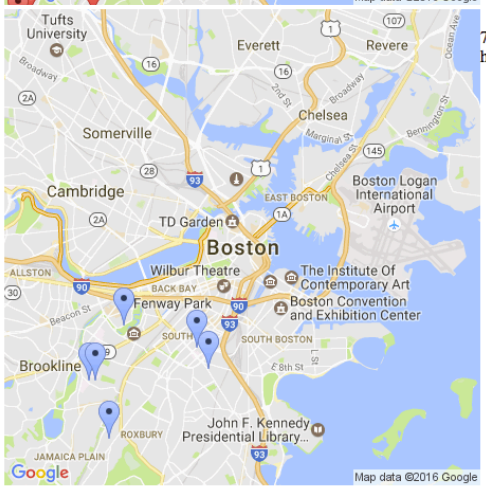
The above figure has zip codes as the X axis, the left Y axis as average property value, and the right Y axis as the number of hospitals. The blue dots represent (zip code, number of hospitals) and the red dots represent (zip code, average property value). This figure was created using the d3js library and used the intersection (by zip code) of the datasets average property value and number of hospitals.

Preexisting hospitals

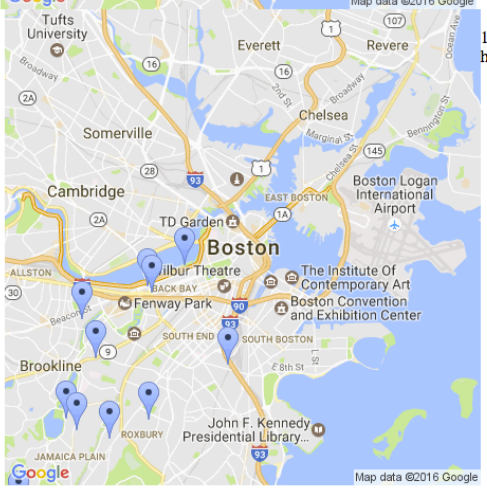




3 Newly added hospitals



7 Newly added hospitals



13 Newly added hospitals

This visual shows preexisting hospitals currently located in and around Boston identifiable with red markers. The newly added hospitals ranging from 3 to 13 can be seen with blue markers.

Conclusions:

The correlation values were reasonably calculated, but may not have much real meaning. There are too many outside factors that can influence any of the statistics we calculated that weren't accounted for. E.g., we can at least say that we did observe high numbers of hospitals and fast food establishments in areas of high property value, but this doesn't fully describe why this is the case; only that it has been observed.

The clustering → k-means paradigm may not output realistic results (lat/long may be in an unreasonable area to build a new hospital), but is a good model to build on for future iterations of algorithms that seek to fully solve resource distribution problems.

Future Work:

The crime incident report dataset had a lot more detail than we took advantage of. Our calculations weighed all the crimes as the same, however not all crimes are equal. In the future, we should assign some score to the crime based on its severity (e.g. a murder would be more

egregious than petty theft). This will lead to more accurate results when using the data from the crime incident reports.

Zip codes may not be a great way of defining neighborhoods. It's possible that somehow creating better / more specific boundaries may lead to more realistic results.

Find more resources to analyze; not only new hospital locations.