

Project Proposal

Sean Smith, Viviana Yee
{swsmith, vivyee}@bu.edu

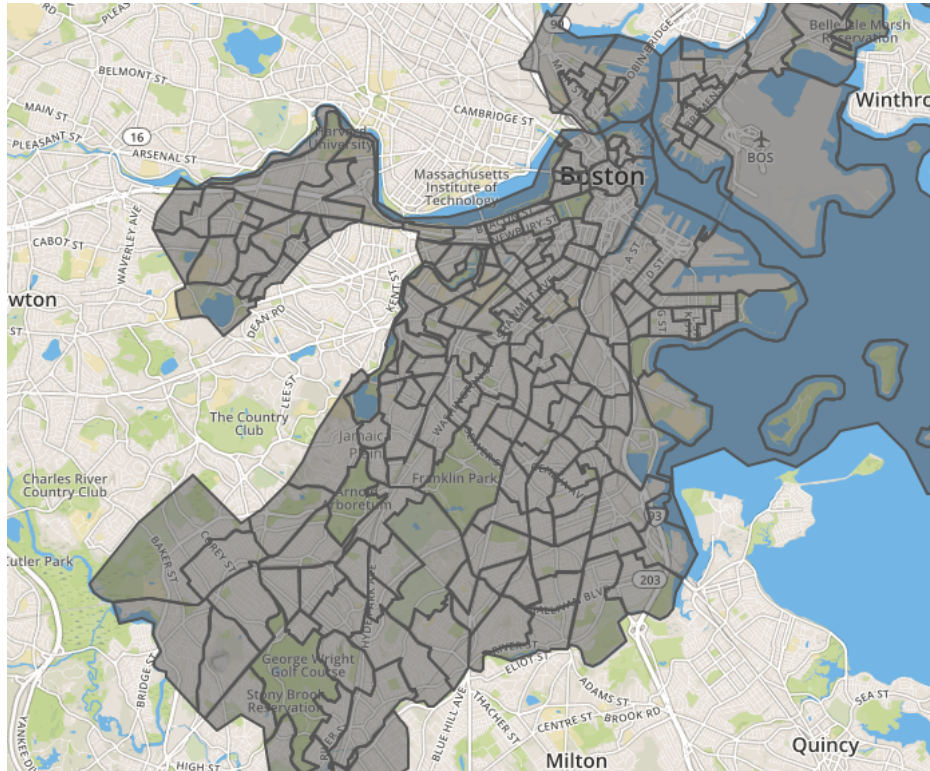
15 March 2017

1 Problem Definition

The city of Boston maintains an open data initiative that provides datasets on lots of different city services <https://data.cityofboston.gov/>. We both have experience taking large amounts of data and making conclusions based on trends and correlations between different datasets. Based on Viviana's experience in Andrei Lapet's Data Mechanics Course and Sean's experience with New York City Taxi data and geographical zone information we've decided to look at Boston obesity data and locations of nutrition centers and the distance between them. Specifically we want to model MBTA transportation options from different neighborhoods (for which we're able to calculate the obesity rate) to nutrition centers. We aim to model the distance in order to answer our hypothesis "Do neighborhoods further from nutrition centers suffer from greater obesity rates?".

2 Data Sources

First we needed to define the geographic areas that we'll be modeling, to do so we've downloaded the 2010 census tracts dataset. This breaks Boston up into unique tracts that each have approximately the same number of residents. You can see an example below:



Next, we took Obesity data from the CDC, Center for Disease Control, and Nutrition Program data from the city of Boston. We filtered these two data sets for longitude and latitude data, so it is valuable for our problem.

Finally, we called the MBTA API endpoints for MBTA stop data. We did this by getting routes for the bus and T. From that, we individually got all stops for each route because there is no API call for all stops in the MBTA system.

By modeling our data with propositional logic we hope to answer the following questions.

- Do neighborhoods further from nutrition centers suffer from greater obesity rates?
- Do neighborhoods with more than one nutrition center have lower rates of obesity?
- Does the obesity rate in surrounding neighborhoods change your neighborhood's access to nutrition centers?

3 Methodology

We're getting datasets from Boston Open Data portal using python and their API's. We're then solving for satisfiability of our WFF's using Z3 <https://github.com/Z3Prover/z3>.

We're doing the following experiments:

1. We're solving for the satisfiability that a coordinate pair is contained in a certain neighborhood. This uses the point in polygon algorithm that Sean developed for the Multi Party Computation class.
2. Find shortest path to all nutrition programs and find shortest out of that
3. Pick a neighborhood with a nutrition program that is geographically closest, and find shortest path to that
4. Combine (find all neighborhoods within certain distance and calculate shortest path for programs within the radius)
5. Find shortest path from every neighborhood to every nutrition center (depends on running time)

4 Progress so far

- We gathered all the obesity and nutrition program data and filtered them for ones that contained useful (longitude and latitude) information.
- We transformed Sean's geographical polygon zone function into a function that uses z3 instead.
- We found which neighborhood each nutritional program and obesity data point were in.
- We got all mbta data needed and also determined the neighborhoods of each stop.

For the rest of the term, we will be working on the shortest path part of the project.

We've setup a github repo so that we can collaborate on code here https://github.com/sean-smith/boston_health (Note that due to size constraints on github we won't be hosting the larger datasets there).