Optimizing Polling Locations Based on Public Transportations

Yueyan Chen (cyyan) Zirui Liu (liuzirui)
Young Jun Choi (yjunchoi) Yuchen Zhang (yzhang71)
CS 591 Data Mechanics – Fall 2017

Introduction

According to Pew Research Center, over 40 percent of the U.S. citizens did not participate in the 2016 presidential election, remaining in the lower bounds compared to other comparable democracies.[1] This is problematic in representative democracies like the U.S., where the political system carries a fundamental assumption of political equality and participation, yet from findings over time made by political scientists, "inequality of representation and influence is not randomly distributed, but are systematically biased in favor of more privileged citizens."[2] Knowing this systematic concern, many states have tried various solutions to fix the low voter turnout problem, yet the situation has not improved in the last two decades. As one of the factors that attracted attention to boost the turnout rate, many studies in the past have examined the relationship between public transit and the voter turnout. The widespread conclusions were pessimistic about the effect, in comparison to the fundamental issues of enthusiasm.[3] Nevertheless, in this research, we wanted to test to see if the public transportations were efficiently and optimally located for the purpose of voting in the first place. With Boston's city-wide public transportation system, we decided that Boston would be a good proxy to test apply this finding. In our project, we reconfigured 255 optimal polling locations across the city, assuming the residents' prevalent use of the Massachusetts Bay Transportation Authority subways and buses.[4]

Datasets

We collected 4 data sets to reconfigure 255 optimal polling locations in 22 wards across Boston.

- Wards: Geospatial data for wards in Boston (https://data.boston.gov/dataset/wards)
- Polling Locations: Set of polling location coordinates in city of Boston (https://data.boston.gov/dataset/polling-locations)
- Bus Stops: Set of bus stop coordinates in city of Boston (http://datamechanics.io/data/wuhaoyu_yiran123/MBTA_Bus_Stops.geojson)
- MBTA: Set of MBTA T station coordinates in city of Boston (http://erikdemaine.org/maps/mbta/mbta.yaml)

After retrieving each dataset and storing them in our database (MongoDB), we divided polling locations, bus stops, and MBTA subway stations into 22 wards based on their coordinates.

[1] Drew Desilver, "U.S. Trails Most Developed Countries in Voter Turnout," *Pew Research Center*, May 15, 2017.
[2] Arend Lijphart, "Unequal Participation: Democracy's unresolved dilemma," *American Political Science Association*, March 1997, Vol. 91, No. 1
[3] Sam Sturgis, "Could Free Public Transportation Get Americans to Voting Booths?" *CityLab*, November 3, 2014.
[4] Nick Wallace "The Best Cities for Public Transportation." *SmartAsset*, July 6, 2017.

(pollingLocation.py, bus_by_ward.py, MBTA_by_ward.py) Throughout our research, we used these new assembled datasets to optimize polling locations in Boston.

Methodology

We used k-means algorithm for optimization, and sampled 10000 Boston voters to compare results from each optimization.

1. A k-means algorithm (optByPublicT.py; optByBusstop.py; optByMBTA.py)

   We used k-means algorithm to find the optimal polling locations based on public transportations, bus stops, and public transit. In each file, we used K-means algorithm to find 255 optimal polling locations in each ward. For optByPublicT.py, we first merged the data sets for bus stops and MBTA, and computed a k-means algorithm with both bus stops data set and MBTA T station data set. For the other two files, we computed a k-means algorithm with each data set. Three files return a different list of polling locations in each ward.

2. Statistical Analysis with Sampling and Inference

   Because it is difficult to tell which optimization method is the best without scoring or evaluating locations, we performed statistical analysis with four different lists of polling locations. We randomized 10,000 addresses in Boston for voters' addresses, instead of using every voter's address in Boston to compare polling locations we optimized. By calculating Euclidean distance between randomized voter's address and the nearest polling location, we determined which optimization method provides the highest accessibility to Boston voters. Throughout the distribution of distance between voters and the polling location, scoringLocation.py returns the result of statistical analysis in 95% confidence interval.

Although our optimization improves the accessibility to the polling location, our optimization results are not perfect solution to improve voter turnout. First, a k-means algorithm does not consider other factors which influence voter turnout. It does not guarantee any usability as polling locations. Moreover, because some polling locations share same coordinates, a k-means algorithm scattered them to separate locations. Therefore, it is hard to determine our optimization results are more accessible than original polling locations.

Visualization

In the scatterplot below (Figure 1), we used blue points to denote “original polling locations” and we used red points to denote “polling locations optimized with public transportation”. Since there are so many points in a single scatterplot, it is difficult for us to compare the difference between these two polling locations and determine which one is better. Same situation happens when we try to compare the original polling locations that are the polling locations optimized by MBTA, and polling locations optimized by bus stops. To solve this problem, we decided to visualize on the map separately and compare results in one table.

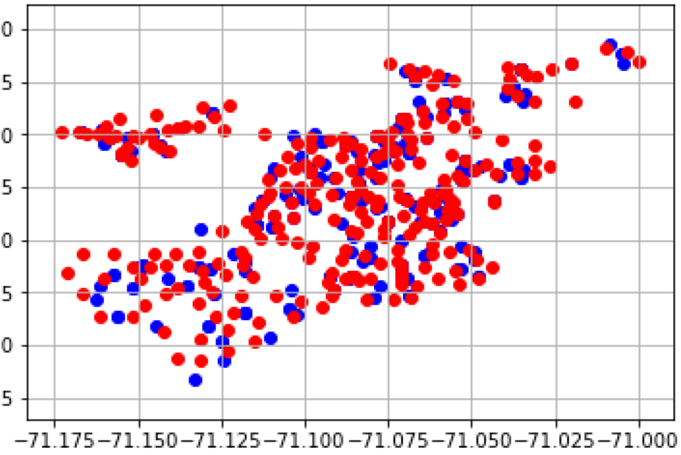
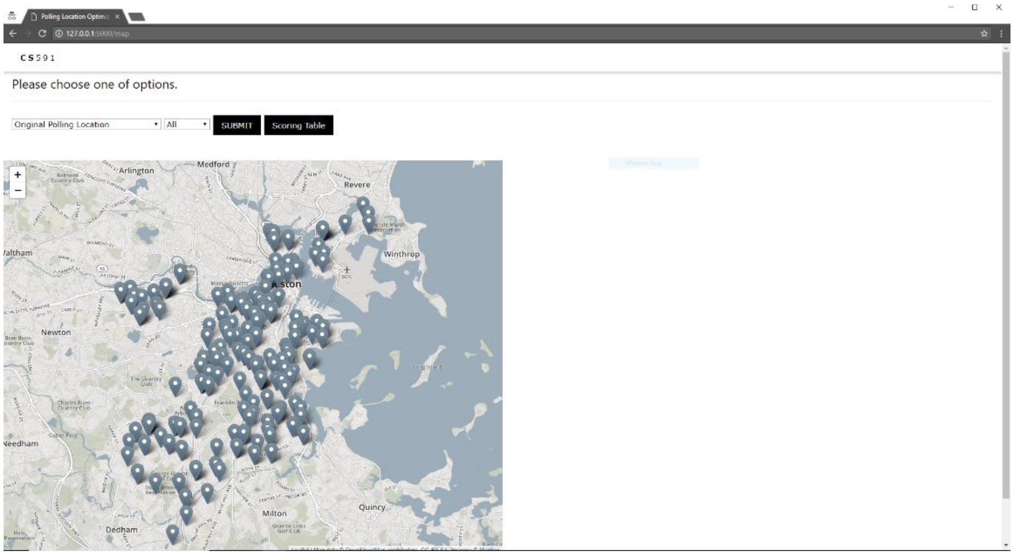


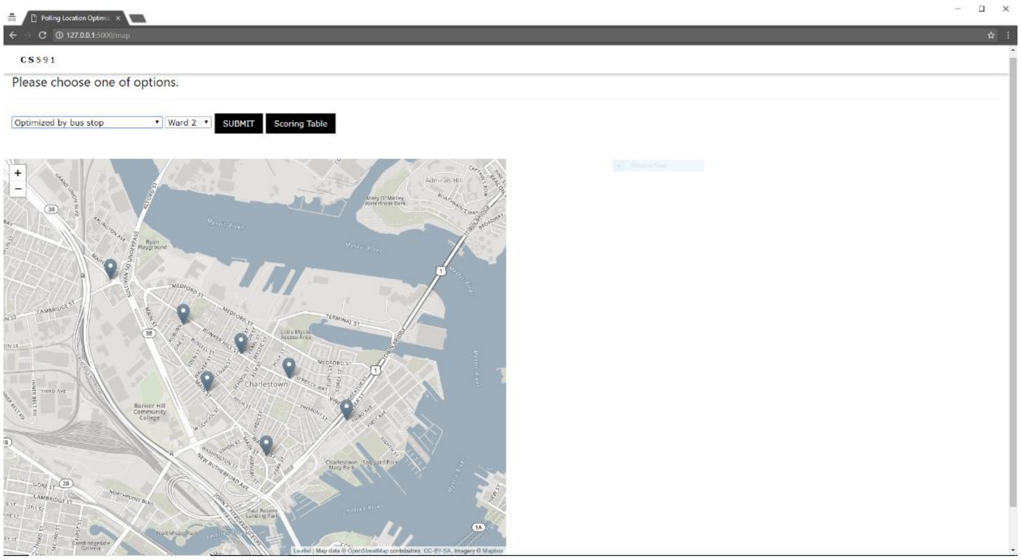
Figure 1: *Scatterplot for original polling locations (blue) and polling locations optimized with public transportations (red).*

In order to visualize our result more clearly, we created an interactive map (Figure 2, 3). In this map, we have two dropdown boxes: one for result, and the other for ward. First dropdown box has four choices: original polling locations, polling locations optimized by bus stops, polling locations optimized by MBTA, and polling locations optimized by public transportations. For each of these four choices, we can also explore the polling location for every ward or explore one specific ward in more details. For example, we choose "polling locations optimized by bus stops" and "Ward 2" in dropdown boxes, the map will show all the polling locations optimized by bus stops (a k-means algorithm) in ward 2 only. Therefore, users can use this map to filter out all the other polling locations they are not interested in and focus on specific ward.



Please choose one of options.

Figure 2: *Interactive map screenshot for original polling locations in every ward.*



CS591

Please choose one of options.

Optimized by bus stop ▾ Ward 2 ▾

Figure 3: *Interactive map screenshot for optimized polling locations in ward 2.*

In this table below (Table 1), we tried to compare the mean (the distance between the voter locations we randomly generated and the polling locations), the standard deviation, and 95% confidence interval of original polling locations and the polling locations optimized by bus stops, MBTA T station, and public transportations (bus stops and MBTA T station combined).

From the data, we concluded that the original locations are actually, optimal polling locations, yet with slight differences between the accessibility optimizing with bus stops versus the MBTA T stops. The descriptions of our finding below will explain the reason and how the differences are reconciled in the overall optimization with both bus stops and the MBTA T stops.

We did not find significant differences between the original polling locations and the locations we have optimized. However, all three kinds of the optimized polling locations are slightly more accessible than the original polling locations. Moreover, among three kinds of optimized polling locations, optimization with bus stops is better than optimization with MBTA, as we find a greater number of bus stops in our data set compared to that of the MBTA T stations. This can also be explained by the confidence interval. This means that the wider the confidence interval, the smaller will the sample size be, and vice versa. Therefore, the confidence interval for “optimization with MBTA” is wider than that for “optimization with bus stops,” because the sample size of the bus stops data sets is much larger. (The width of the confidence interval for “optimization with bus stops” is 2.484, and the width of the confidence interval for “optimization with MBTA” is 2.76.) With this in mind, we can also understand why the data for “optimization with public transportation” is so similar to the data for “optimization with bus stops” because when we merge the bus stop data set and the MBTA T station data set, most portion of the data set are data related to bus stops, but not MBTA T station.

Polling Locations	Average (Miles)	STD (Miles)	95% Lower Tail Confidence Interval (Miles)	95% Upper Tail Confidence Interval (Miles)
Original Polling Locations	1.035	0.828	0.069	2.967
Optimization with Bus Stops	0.828	0.69	0.069	2.553
Optimization with MBTA T station	0.966	0.828	0.069	2.829
Optimization with Public Transportations	0.828	0.69	0.069	2.484

Table 1: *Scoring table to compare original polling locations, polling locations optimized by bus stops, polling locations optimized by MBTA, and polling locations optimized by public transportations.*

Conclusion

As the table (Table 1) shows, each result has an improvement over original polling locations. Polling locations optimized with bus stops and with public transportations (bus stops and public transit combined) are more accessible than polling locations optimized with MBTA T stations because the number of bus stops is greater than MBTA T stations. However, in this research, we found that some original polling locations are in the same building but each location is assigned for different people. By running a k-means algorithm we have more markers on the map (Figure 1, 2), which improves the accessible score of each result of optimization. Therefore, we found that the county commission already chose enough accessible polling locations for Boston voters. Still, the work presented here will be developed more for future studies of America's voter turnout problem with considering more factors to increase voter turnout rates.

Future Work

Our next step to improve this research is routing between the nearest polling locations and randomly generated voters' addresses. In this research, we did not consider how people would get to the designated polling location. Therefore, with considering the route to polling locations, we will be able to determine which polling locations are more accessible to Boston voters. Furthermore, since we ran a k-means algorithm to optimize on public transportations, we did not consider building usage, openness to public, and so on. Therefore, our optimal polling locations might not be ideal polling locations in reality. For example, some polling locations might be too small or private properties. Therefore, we will develop the scoring methods to evaluate suitability as a polling location.

References

Arend Lijphart, "Unequal Participation: Democracy's unresolved dilemma," *American Political Science Association*, March 1997, Vol. 91, No. 1

Drew Desilver, "U.S. Trails Most Developed Countries in Voter Turnout," *Pew Research Center*, May 15, 2017.

Nick Wallace "The Best Cities for Public Transportation." *SmartAsset*, July 6, 2017.

Sam Sturgis, "Could Free Public Transportation Get Americans to Voting Booths?" *CityLab*, November 3, 2014.