The Effect of Land Use on "Walkability" and Within Boston's Bicycle Ridesharing Network

Abstract

Public transportation within the Greater Boston metropolis serves as a priceless asset to its citizens. However, automotive congestion and urban terraforming threaten a successful transition to multi-mode, sustainable public transportation – specifically threatening ridership with Boston's "BlueBike" bicycle ridesharing program. We test the efficacy of using isochrone polygons to represent "walkable" area within 4 Boston Municipalities and cluster bicycle ridesharing stations based on isochrone land use to discern possible interconnectivity within the ridesharing station network, and test overall robustness of the model. Results show walkable areas contain primarily commercial, residential, and industrial land (which skews the overall robustness of the model), where commercial and transportation land cluster in tandem.

Introduction

Today, with the onset of climate change, rising urban population density, health hazards from pollution, and urban sprawl, cities are slowly losing transportation interconnectivity and communal idiosyncrasies to automotive-focused urban design. Specifically, in Boston, Massachusetts, there exists a prominent dearth of walking- and cycling-friendly space around the entire city. Understanding the factors that influence the public's choice to utilize alternative transportation is of paramount importance to civic health in Boston, and in cities throughout the US. The nexus between land use/walkability[1] and land use/cycling[2] have already been extensively analyzed, showing that commercially friendly land causes the sharpest response in behavior from citizens. Municipal governments within the United States[3] and United Kingdom[4] have already conducted applied analyses on factors inhibiting cycling throughout urban areas, finding that the option to walk, age, incompatible transportation infrastructure, and quick driving commute distances decrease cycling. Walking-based urban isochrone analysis, although relatively new, has prominence within the literature,[5] and k-means clustering of cycling behavior has become a recent benchmark component of spatial analysis.[6]

---

[1] Duncan et. al. "Relationships of Land Use Mix with Walking for Transport: Do Land Uses and Geographical Scale Matter?"
[2] Yuchen et. al. "Land Use Effects on Bicycle Ridership: A Framework for State Planning Agencies."
[3] Federal Highway Administration. "Case Study #1"
[4] Victoria Transport Policy Institute. "Land Use Impacts on Transport."
[5] Dovey et. al. "Isochrone Mapping of Urban Transport: Car-dependency, Mode Choice, and Design Research."
[6] Xu et. al. "Station Segmentation with an Improved K-Means Algorithm for Hangzhou Public Bicycle System."

Research Question

This study's query provides the most suitable vector for the intended research efforts: How do land use and/or zoning classifications within Greater Boston coincide with bicycle ridesharing behavior, specifically within the municipalities of Boston (proper), Cambridge, Somerville and Brookline? The more prudent question would be: can zoning classifications affect behavior? Additionally, as a relevant subquery: Is the model we choose to use impartial, empirically robust, and sufficiently malleable for future research? These research questions are unique within urban studies due to the new center of connectivity they attempt to deliver upon. The intersection of land use and cycling, as well as land use and "Walkability", have primarily been the domain of city planners (and to a lesser degree health researchers), and the link between "Walkability" and cycling has been the domain of federal research (at least within the US and UK). These questions bridge the 3 domains in a new, interdisciplinary model.

Data Source & Collection

BlueBike station data is extracted from official BlueBike system data.[7] Neighborhood boundary data for Boston is extracted from Analyze Boston[8], boundary data for Cambridge is extracted from the City of Cambridge[9], boundary data for Somerville is extracted from the City of Somerville,[10] and boundary data for Brookline is extracted from MassGis[11] The 29 census tracts used to construct the Dorchester, Boston neighborhood are extracted from the US Census Bureau.[12] Finally, land use/zoning data for all municipalities is extracted from MassGis.[13]

The process of data cleaning is as follows: First, town data and neighborhood boundary data are imported into CARTO. Next, BlueBike station data is filtered into 252 stations based on unique origin and destination locations in Pandas (Python) and exported to CARTO for visualization. Next, land use

[7] Motivate International Inc. "System Data."
[8] Analyze Boston. "Boston Neighborhoods."
[9] City of Cambridge, MA. "CDD Neighborhoods."
[10] Johnson, Keith. "Neighborhoods."
[11] MassGIS (Bureau of Geographic Information). "MassGIS Data: Community Boundaries (Towns)."
[12] United States Census Bureau. "Cartographic Boundary Shapefiles - Census Tracts"
[13] MassGIS (Bureau of Geographic Information). "MassGIS Data: Land Use (2005)."

data is imported into Pandas (Python), and 31 land uses are assigned 1 of 11 customized land use classifications.[14] Next, Census Tract data is imported into CARTO and manipulated with PostGIS SQL (in CARTO) to extract the 29 census tracts necessary to construct the neighborhood of Dorchester, Boston. Finally, Pandas (Python) is used to arrange all cleaned data aforementioned into multiple dataframes as necessary to construct summary statistics. and prepare data for analysis in a K-Means Clustering matrix, which is then cleaning in Pandas and exported to CARTO for visualization.

Methods, Measurements & Design

This study is designed in 3 parts: Rudimentary land use analysis and data exploration, sequestered "walkability" isochrone analysis, and finally K-Means Clustering analysis. For rudimentary land use analysis, the 252 unique BlueBike stations are displayed atop 11,619 zones for (1) qualitative geospatial analysis of 11 customized land use types and (2) the construction of a LUDI score at the neighborhood level (for all 59 neighborhoods) based on the area of 7 customized land use categories created from the original 11 using Pandas (python).[15] Area is calculated in hectares for all reported measurements throughout this study. Summary statistics are then calculated for LUDI scores and land use. The equation used to calculate the LUDI is as follows:

$$LUDI = 1 - \frac{\left|\frac{I}{T} - \frac{1}{6}\right| + \left|\frac{TR}{T} - \frac{1}{6}\right| + \left|\frac{R}{T} - \frac{1}{6}\right| + \left|\frac{IN}{T} - \frac{1}{6}\right| + \left|\frac{C}{T} - \frac{1}{6}\right| + \left|\frac{N}{T} - \frac{1}{6}\right| + \left|\frac{OS}{T} - \frac{1}{6}\right|}{\frac{5}{3}}$$

Where I = Industrial land, TR = transportation land, R = residential land, IN = Institutional land, C = Commercial land, N = Natural land, and OS = Open space.[16] Contrary to Rajamani et. al., natural land was included due to the possibility of erratic zoning history (for example, some open spaces are

[14] See appendix for classification of each original land use into customized land use.
[15] The 4 zoning categories "Public Services, Recreation, Agriculture, and Other" are removed due to relative underrepresentation and irrelevance to the scope of study.
[16] Rajamani, J., C. Bhat, S. Handy, G. Knaap, and Y. Song. "Assessing Impact of Urban Form Measures on Nonwork Trip Mode Choice after Controlling for Demographic and Level-of-service Effects."

zoned as national, state, or city parks). For operationalization, the output of the LUDI equation is calculated in PostGIS SQL, and then migrated to Pandas (python) for further analysis.

For sequestered "walkability" isochrone analysis, isochrones are creating using the OpenRouteSource API (based on OpenStreetMap) for 5, 10, and 15 minutes of walk time from their centroids for each of the 252 unique BlueBike stations.[17] These isochrones are then visualized in CARTO, and land use composition of each isochrone based on the 7 customized zones aforementioned from only the 5-minute isochrones are extracted and placed in PostGIS datasets. The isochrone centroid was chosen based on the following concept: One could walk from any point in the isochrone to the centroid in 5 minutes or less, or one could walk from the centroid to any point in the isochrone in 5 minutes or less. Summary statistics for 5-minute isochrones are then calculated.[18]

We define an isochrone as a convex-hull polygon emerging from a centroid outward at a fixed distance in every direction based on *x* number of minutes one can travel on foot, according to estimates from OpenStreetMap. These polygons are created using the OpenRouteSource API (based on OpenStreetMap). Summary statistics for these polygons are calculated in Pandas (python), and the 2 OLS multi-variate regressions calculated to determine correlation between commercial and residential land within the 5-minute isochrones are constructed/operationalized in R and are defined as follows: [19]

$$area_{com} = \beta_0 + \beta_1 area_{res} + \beta_2 area_{ind} + \beta_3 area_{inst} + \beta_4 area_{nat} + \beta_5 area_{os} + \beta_6 area_{trans} + \epsilon$$

$$area_{res} = \beta_0 + \beta_1 area_{com} + \beta_2 area_{ind} + \beta_3 area_{inst} + \beta_4 area_{nat} + \beta_5 area_{os} + \beta_6 area_{trans} + \epsilon$$

For K-Means Clustering analysis, the aforementioned PostGIS dataset for 5-minute isochrones is used as the subject of the clustering analysis, in which the area corresponding to the 7 customized zones (identical to those used in the LUDI calculation above) contained within each isochrone of the PostGIS dataset are used to weight each cluster. K-Means clustering is performed for clustering coefficients k=2 through k=8, and k=3 is chosen as the best fit. The analysis is operationalized using Scikit-Learn

[17] See Appendix for larger photograph of all 3 isochrone types overlapping.
[18] See Appendix for statistics regarding the sum of all land use types within the 5-minute walkability isochrones.
[19] See Appendix for full results on both regressions

(python), and the presence of each of the 7 land uses within all 252 stations across the 3 clusters are visualized with Seaborn (python).

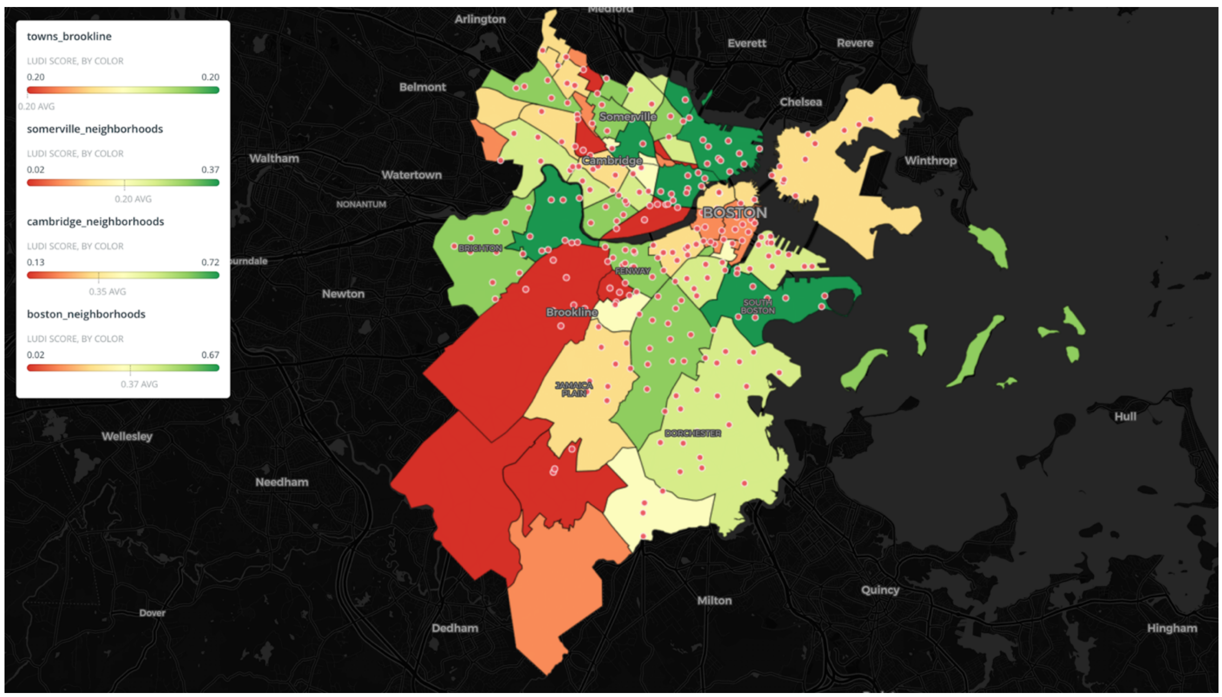
Results

Basic visualization show most of Greater Boston is dominated by residential, commercial, and industrial land use (a common theme). Commercial land dominates most of Downtown Boston, while Boston's suburbs are dominated by residential land. Cambridge has the greatest diversity of land use between the 4 municipalities, and both Somerville and Brookline are almost entirely characterized by residential land use. The following map plots the 11,619 land use data points and 252 BlueBike stations (Figure 1):

The following table represents summary statistics for all 11 customized land use categories throughout all 4 municipalities (Figure 2):

	land_use	sum_area	mean_area	median_area	min_area	max_area	std_area	percentage_area	count
0	Agriculture	47.577914	1.189448	0.516355	3.687550e-02	15.92700	2.560560	0.269522	40
1	Business/Commercial	1794.315020	1.209922	0.214697	2.323320e-05	198.89600	6.048092	10.164552	1483
2	Industrial	812.853704	1.963415	0.408366	3.914810e-03	53.90370	4.722760	4.604706	414
3	Institutional	1866.697980	0.942301	0.110530	1.836520e-05	111.25600	4.452085	10.574591	1981
4	Natural	2352.603671	2.673413	0.458553	1.680700e-07	289.27900	13.676682	13.327181	880
5	Open Space	278.716301	1.459248	0.525240	9.089840e-03	28.66270	2.893823	1.578890	191
6	Other	67.226755	1.244940	0.587554	2.654120e-03	9.00903	1.812276	0.380830	54
7	Public Services	524.660191	10.089619	1.469165	8.282860e-03	95.45620	21.404183	2.972129	52
8	Recreation	1047.259392	2.727238	1.022045	1.676270e-05	68.04100	6.484041	5.932583	384
9	Residential	7325.784345	1.205097	0.121557	7.431090e-06	338.07700	9.866598	41.499577	6079
10	Transportation	1534.977235	25.163561	1.747230	1.206240e-03	639.08000	92.929394	8.695438	61

Summary Statistics for land use show similar results to basic visualization and also introduce another motif that appears throughout our analysis: Transportation is also highly represented throughout Boston, as it remains the 4th most prominent land use type by summed land use. The highest mean, median, and maximum values belong to transportation land use. Residential land represents the highest quantity of data points, followed by Institutional and Business/Commercial. The most volatile of the land use types is Public Services land, possible due to its very high range and relatively low quantity. Residential land represents the greatest percentage of land use, followed by Natural and Institutional land. Next, we present the visualization (Figure 3) and summary statistics (Figure 4) of the LUDI at the neighborhood level:

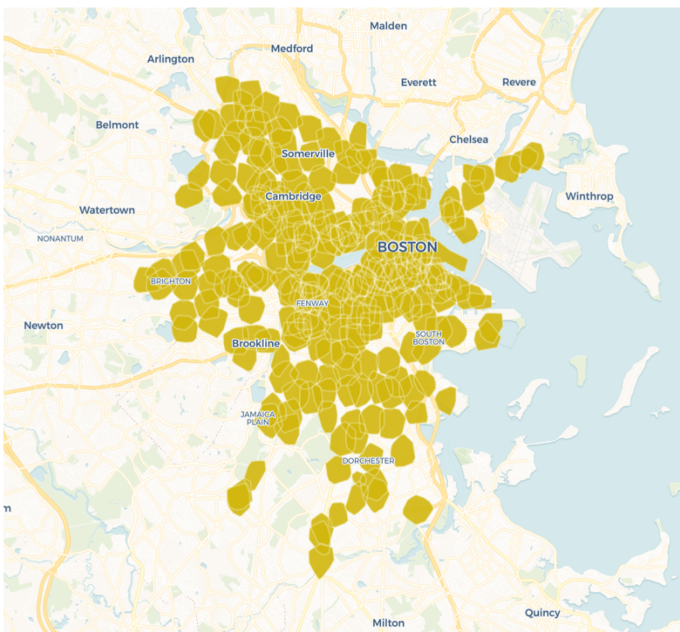


	ludi
count	26.000000
mean	0.367562
std	0.149254
min	0.023722
25%	0.317625
50%	0.364147
75%	0.461498
max	0.665705

	ludi
count	13.000000
mean	0.352224
std	0.172767
min	0.133757
25%	0.244367
50%	0.316763
75%	0.449116
max	0.724774

	ludi
count	19.000000
mean	0.196037
std	0.092085
min	0.015988
25%	0.119184
50%	0.210584
75%	0.247731
max	0.373437

From left to right, summary statistics (Figure 4) represent Boston, Cambridge, and Somerville. The LUDI for Brookline is 0.3002. Again, commercial dominance of downtown Boston begets a low LUDI, while the suburban outskirts of Boston show very low LUDI. Jamaica Plains and other large neighborhoods closer to downtown show higher LUDI scores, with South Boston displaying the highest. Again, Cambridge displays high land use diversity, as represented by the second highest average LUDI of Greater Boston, and both Somerville and Brookline are primarily dominated by residential land, with Somerville showing the greatest variance of LUDI throughout Greater Boston. Summary statistics of Greater Boston's LUDI scores echo these results. Boston holds the highest mean and maximum LUDI values, while Cambridge holds the Somerville holds the lowest minimum value. Cambridge also holds the highest variance of all LUDI scores. Next, visualization of all 252 5-minute walkability isochrones and total percentage statistics, and complete summary statistics (Figure 5, 6, 7) are displayed below.

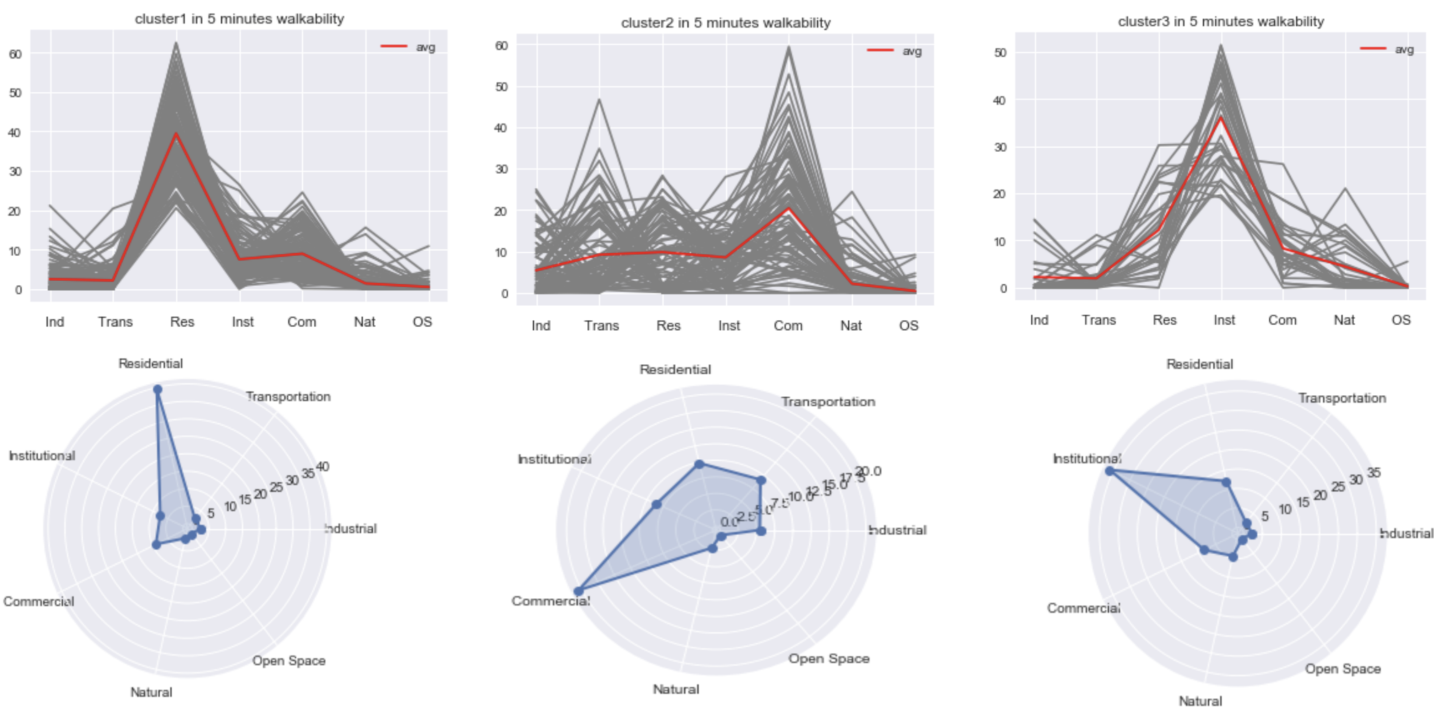


	measure	result
8	p_ind	5.617882
9	p_trans	7.545472
10	p_res	38.746194
11	p_inst	18.137989
12	p_com	20.676880
13	p_nat	3.323416
14	p_os	0.749921
15	p_total	94.797754

	area_ind	area_trans	area_res	area_inst	area_com	area_nat	area_os	area_hec
count	252.000000	252.000000	252.000000	252.000000	252.000000	252.000000	252.000000	252.000000
mean	3.549608	4.767538	24.481431	11.460323	13.064499	2.099870	0.473831	63.184091
std	5.119264	7.224827	17.161600	11.131710	10.827654	3.897147	1.331930	11.003827
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	17.401205
25%	0.115128	0.374901	7.741802	3.907499	5.258303	0.000000	0.000000	59.029345
50%	1.369541	1.709644	23.324217	8.049854	10.085753	0.254540	0.000000	66.181880
75%	4.999472	5.687542	39.189276	14.279001	17.148459	2.332390	0.384928	70.129853
max	24.984115	46.669509	62.505084	51.426326	59.410564	24.413074	10.916163	80.507805

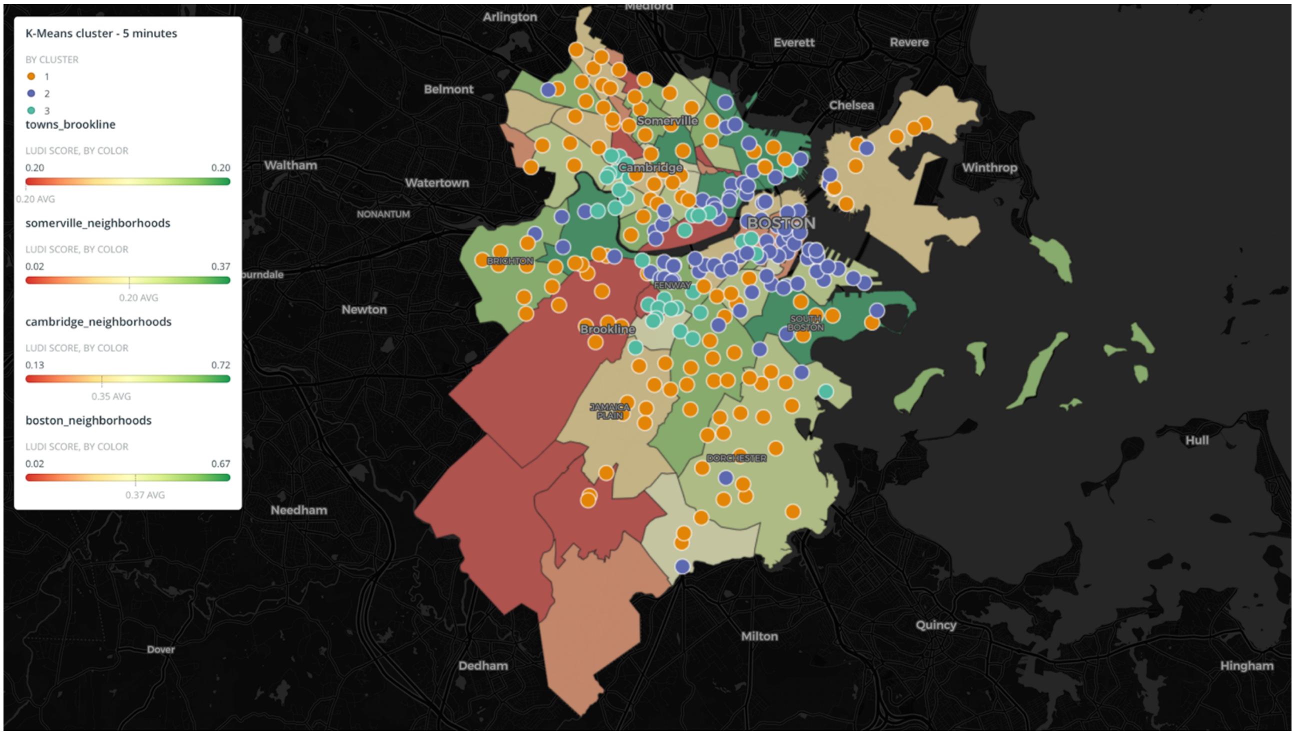
We observe that isochrones often overlap and are quite misshapen in certain areas. Regardless, 94% of land across all isochrones are represented by the 7 customized land use types, meaning that the customized land use choices represent most land within walkable area around BlueBike stations. However, total area is not balanced/equivalent across all isochrones. Residential land holds the highest maximum and median values. Residential and institutional land have the highest variance. Additionally, residential and commercial land represent the greatest total area of all isochrones. Finally, the regression computed to determine the possible correlation of commercial and residential land use show that for every 1 hectare increase of residential land, 0.5 hectares of commercial land is lost, and conversely, for every 1 hectare increase of commercial land, 0.659 hectares of residential land is lost.

The following graphs numerically represent the specific land use concentration of each k-means cluster (Figures 8A-8F):



Cluster 1 is the most uniform cluster, primarily focused on residential land, where commercial land displays weak coexistence. Cluster 2 is the most dispersed cluster, and is primarily focused on commercial land, with transportation land displaying a strong coexistence, along with natural and residential land. Cluster 3 is the most erratic (non-uniform) cluster, primarily focused on institutional land, with equal second order representation by commercial and residential land, and equal third order representation by natural and transportation land.

Finally, the visual results of the K-Means Clustering (with k=3) atop neighborhood-level LUDI is plotted below (Figure 9):



Echoing previous results, the 3 clusters are concentrated as follows: Cluster 1 is the loosest concentration and the largest group of clustered stations, scattered among both the residential and diversified land use areas of Greater Boston, as well as parts of East Boston and Brookline. Cluster 2 is the tightest and highest density cluster, focused primarily on Downtown Boston, Seaport, and the Charles River waterfront areas within Cambridge boundaries. Cluster 3 is the smallest group and most dispersed, focused on the boundary between urban and suburban areas within Boston and Cambridge.

Conclusion, Policy Implications & Future Steps

In conclusion, rudimentary land use & LUDI analysis indicate that Commercial and Residential land use dominate the majority of land use across Greater Boston, especially in Somerville, and the modified LUDI equation proved effective at faithfully capturing primary land use within the study area. We see that residential and commercial land dominates greater Boston, but k-means clustering elucidates the common presence of transportation and commercial land – a crucial component to public transportation interconnectivity. However, transportation land within walkable areas is far underrepresented relative to total land use. Next, walkability isochrones proved faithful, capturing 94% of all 11 land use types over all 252 polygons. Regression analysis shows commercial and residential land display a weak negative relationship at miniscule magnitude. Overall, this study finds that land use diversity has little relation with station placement (within walkable area between stations), and this model requires more thorough mathematical and statistical modifications to meet rigorous research standards. In the future, this model could be modified to represent time variant data (current data is time invariant) in order to assess dynamic travel behavior within walkable area surrounding stations. Dynamic travel behavior analysis could prove effective and useful to policymakers and city planners where periodic bicycle rebalancing is concerned.

This study presents 3 main policy implications for municipal/government officials and the public. First, this analysis displays strong relevance to climate change. The City of Boston predicts that between 2050-2100, a large portion of Downtown Boston will experience frequent flooding, and clustering analysis shows that a major portion of ridesharing stations are clustered directly along the waterfront. Second, because overall analysis displays prominence of residential and commercial land, housing policy may be reformed to adapt to the presence of ridesharing stations – possibly providing financial/demographic incentives for regular travel on the BlueBike network. Third (as an indirect effect of the previous 2 results), overall public transportation design and implementation throughout Boston must be reconsidered in order to complement the rise of privately-owned public transportation. Interconnectivity is essential in maintaining the health of Boston's public transportation system, and inter-modal synchronization will be key to increasing the city's reliance on public transportation overall.

References

Analyze Boston. "Boston Neighborhoods." *Data.boston.gov.* Accessed December 9, 2018. https://data.boston.gov/dataset/boston-neighborhoods

City of Cambridge, MA. "CDD Neighborhoods." *Geographic Information System.* Accessed December 9,2018 <https://www.cambridgema.gov/GIS/gisdatadictionary/Boundary/BOUNDARY_CDDNe

ghbohods.aspx>

Cui, Yushen; Mishra, Sabyasachee; and Welch, Timothy F.. "Land Use Effects on Bicycle Ridership: A Framework for State Planning Agencies." *Journal of Transport Geography 41*, no. C (2014): 220-28.

Dovey, Kim, Woodcock, Ian and Pike, Lucinda. "Isochrone Mapping of Urban Transport: Car dependency, Mode-choice and Design Research." *Planning Practice & Research* 32, no. 4 (2017): 402-16.

Duncan, Mitch, J. Winkler, Elisabeth Sugiyama, Takemi Cerin, Ester DuToit, Lorinne Leslie, and Eva Owen. "Relationships of Land Use Mix with Walking for Transport: Do Land Uses and Geographical Scale Matter?" *Journal of Urban Health* 87, no. 5 (2010): 782-95.

Federal Highway Administration. “Case Study 1: Reasons Why Bicycling And Walking Are And Are Not Being Used More Extensively As Travel Modes.” *US Department of Transportation.* Accessed December 9, 2018. < https://safety.fhwa.dot.gov/ped_bike/docs/case1.pdf>

Johnson, Keith. “Neighborhoods.” *Somerville MA – Mayor Joseph A. Curtatone.* Accessed December 9, 2018. < https://data.somervillema.gov/GIS-data/Neighborhoods/n5md-vqta>

MassGIS (Bureau of Geographic Information). “MassGIS Data: Community Boundaries (Towns).” *Massdocs Pilot.* Accessed December 9, 2018. < https://docs.digital.mass.gov/dataset/massgis data-community-boundaries-towns>

MassGIS (Bureau of Geographic Information). “MassGIS Data: Land Use (2005).” *Massdocs Pilot*. < https://docs.digital.mass.gov/dataset/massgis-data-land-use-2005>

Motivate International Inc. “System Data.” *Motivate International Inc.* Accessed December 9, 2018. https://www.bluebikes.com/system-data

Rajamani, J., C. Bhat, S. Handy, G. Knaap, and Y. Song. "Assessing Impact of Urban Form Measures on Nonwork Trip Mode Choice after Controlling for Demographic and Level-of-service Effects." *Transportation Research Record*, no. 1831 (2003): 158-65.

United States Census Bureau. “Cartographic Boundary Shapefiles - Census Tracts.” *US Census Bureau.*

Accessed December 9, 2018. < https://www.census.gov/geo/maps-data/data/cbf/cbf_tracts.html>

Victoria Transport Policy Institute. "Land Use Impacts on Transport." *Victoria Transport Policy Institute.* Accessed December 9, 2018. < http://www.vtpi.org/landtravel.pdf>

Xu, Haitao, Ying, Jing Lin, Fei and Yuan, Yubo. "Station Segmentation with an Improved K-Means Algorithm for Hangzhou Public Bicycle System." *Journal of Software* 8, no. 9 (2013): 2289-296.

Appendix (Please see folder within .zip file for code/programming files)

Sum of area (by land use type) within 5-minute walking distance isochrones (in Hectares)

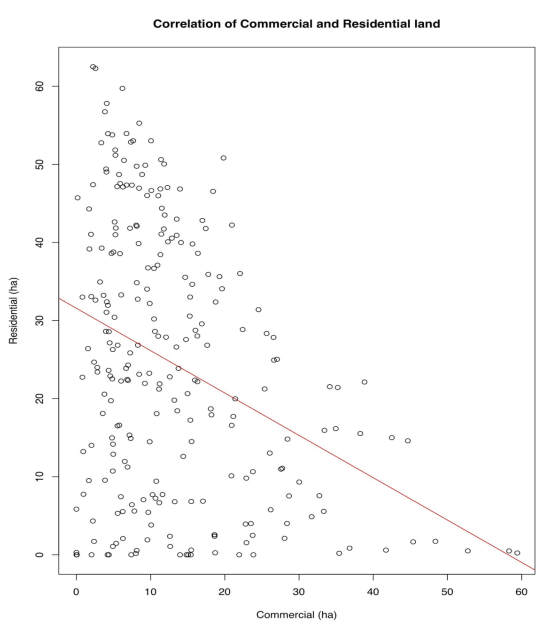
Commercial	Residential	Institutional	Industrial	Transportation	Natural	Open Space
3292.25	6169.32	2888	894.50	1201.42	529.17	199.41

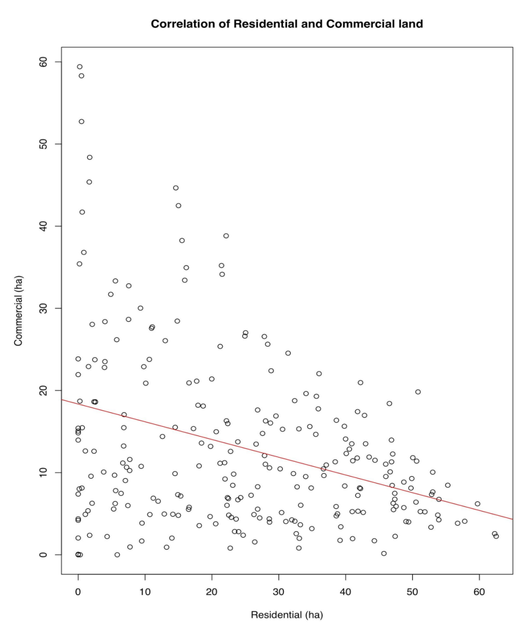
Land Use Classifications Within Summary Statistics Tables

Area_com	*Area_res*	*Area_inst*	*Area_ind*	*Area_trans*	*Area_nat*	*Area_os*
Commercial	Residential	Institutional	Industrial	Transportation	Natural	Open Space

Full Regression Results and graphical output

Regression 1	$\beta_1 = -0.501$	Adusted $R^2 = 0.3469$	$p < 0.000$
Regression 2	$\beta_1 = -0.659$	Adusted $R^2 = 0.6607$	$p < 0.000$

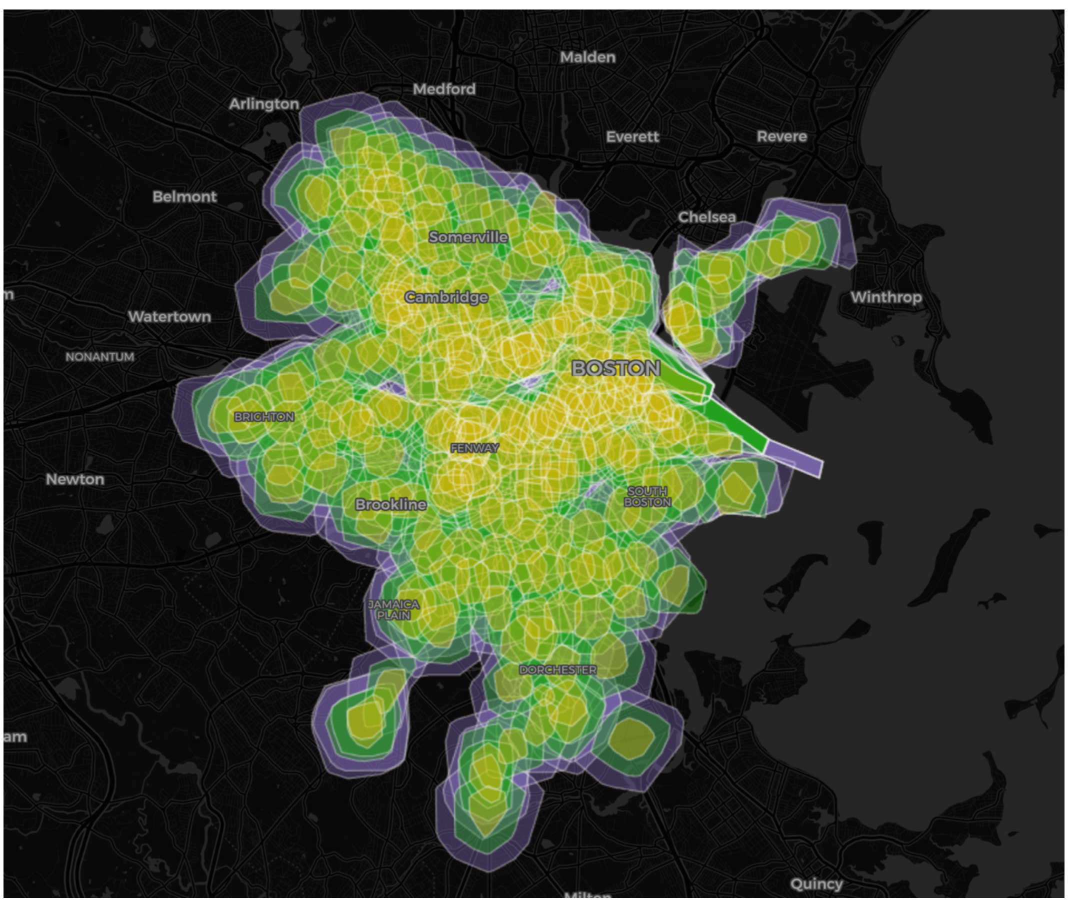




Classification of all 31 land uses into 11 customized land uses from MassGis data

Agriculture – Orchard, Pasture, Cropland, Nursery
Business/Commercial – Commercial
Industrial – Mining, "Industrial"
Institutional – Urban Public/Institutional
Natural – Saltwater Wetland, Non-Forested Wetland, Saltwater Sandy Beach, Forest, Water, Brushland/Successional, Forested Wetland
Open Space – Open Land
Other – Transitional
Public Services – Cemetery, Waste Disposal, Junkyard
Recreation – Golf Course, Water-Based Recreation, Marina, Participation Recreation, Spectator Recreation
Residential – Multi-Family Residential, High Density Residential, Medium Density Residential, Low Density Residential
Transportation - Transportation

All Ischrones Overlaid



[Hyperlink to interactive map comparing station by k-means cluster and LUDI at the neighborhood level]()