Untangling the effects of residential segregation on individual mobility

by

Suma Desu

B.S., University of Vermont (2012)

Submitted to the The Center for Computational Engineering
in partial fulfillment of the requirements for the degree of

Master of Science in Computation for Design and Optimization

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2015 [June 2015]

Author ..... Signature redacted
The Center for Computational Engineering
Jan 30, 2015

Certified by ..... Signature redacted
Marta González
Assistant Professor
Thesis Supervisor

Accepted by ..... Signature redacted
Nicolas Hadjiconstantinou
Chairman, Department Committee on Graduate Theses

Untangling the effects of residential segregation on individual mobility

by

Suma Desu

Submitted to the The Center for Computational Engineering
on Jan 30, 2015, in partial fulfillment of the
requirements for the degree of
Master of Science in Computation for Design and Optimization

Abstract

More than half of today's world population lives in cities and that fraction is steadily growing. Models that accurately capture all segments of the population are necessary in order to design effective policies and new technologies to ensure efficient and stable operations of cities. The current sociology literature has a rich foundation in characterizing the demographics of static population distributions, however, these characterizations fail to account for the reality of dynamic movement. Though there has been recent work in developing models of human mobility, they in turn do not capture demographic differences in the populations of cities.

In this work we present a computational approach to reformulating segregation metrics to incorporate dynamic movement patterns and also quantify the effects of introducing demographics into a mobility model. In coupling two fields that are inherently connected but not established as so, we must very carefully consider our experimental set up. The first part of this work deals with understanding our data and its limitations at fine granularities and explicitly measuring segregation metrics at various scales to design a study that will elucidate meaningful aspects of segregation.

In the second part of this work we reformulate traditional segregation metrics using topological properties of origin destination networks as input. These measures are flexible in considering many locations that individuals visit and therefore more accurately capture the environments of individuals that traditional segregation literature seeks to characterize. We utilize two rank-based mobility models that implicitly incorporate geographic properties of population distributions to understand the effects of residential segregation on mobility patterns and examine the effect of demographic considerations on model accuracy.

In summary, this thesis will focus on synthesizing the rich body of work on static characterizations of socioeconomic structure in cities with dynamic models to better understand different racial segmentations of Boston's population. This work is both an extension to static segregation literature as well as a refinement of current mobility models.

Thesis Supervisor: Marta González
Title: Assistant Professor

Acknowledgments

To my family: I thank you for your tireless support and unceasing kindness. Words cannot express my gratitude, so I hope to live every day in acknowledgement of your grace. It is my greatest fortune to have you as my parents and brother.

To my advisor Marta Gonzalez, thank you for the opportunities you have given me.

To my former advisors Chris Danforth and Peter Dodds, I am humbled to have known you. Thank you for your lasting friendship.

To my friends and lab mates thank you for your inspiring perspectives and your wonderful company.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction and overview

Physicists and sociologists have longstanding interests in identifying the mechanisms by which individual dynamics lead to collective outcomes. In the social domain, tipping point or threshold models provide one useful framework for connecting the actions of individuals to population processes [41, 20, 21]. In the physics domain mobility models can account for individual preferences that result in population level commuting fluxes at multiple scales [19, 43]. The underlying premise of these models stems from the assumption that the actions of individuals are influenced by given characteristics but these characteristics are also influenced by the choices of individuals. Threshold, epidemic, diffusion, and gravity models are part of a general class of behavioral models that capture the feedback effects between micro and macro-level processes. Behavior models, such as models of social interaction, have demonstrated great potential for understanding the dynamics of both residential mobility and segregation by race and ethnicity. In his work [41], Thomas Schelling laid the conceptual groundwork for understanding the relationship between individual preferences and the evolution of neighborhood compositions. He demonstrated that minute racial preferences in individual residential choices could result in aggregate patterns of residential segregation over time. Schelling's ideas provide an account of neighborhood change that formalizes the consequences of prejudice, which has been

documented in social survey data and provides an explanation of manifest patterns of residential segregation.

Residential choices are often the result of careful planning and consideration, thus the importance of individual preference in deciphering patterns of segregation is well established. On the other hand, daily mobility choices, such as choosing a grocery store, are often ephemeral. Although these daily decisions might seem random, individual mobility data has revealed to be highly non-random, governed by simple laws, and greatly predictable [42, 19, 44]. Accordingly we can use individual trajectories as a proxy to understand the environments to which individuals are exposed.

Investigating individual's daily trajectories is a relatively new direction of research which stems from the availability of massive passively generated geo-located datasets. In addition to its non random nature mobility has further been shown that mobility dynamics are subject to geographic constraints [16, 10, 12]. The majority of travel occurs within cities and it has been shown a city's internal structure affects urban scale mobility. Thus, previous work has been done using mobility to estimate the structure of cities by finding hotspots [28] or even finding patterns in the daily encounters of individuals [45]. These studies have focused on geographic distance and social networks, but have not linked the socio economic structure of a city and its resulting mobility patterns. In coupling mobility and social data we can investigate how demographics effect individual mobility.

In this work, we use call detail records (CDRs) generated over multiple months, which are treated and validated, to create average daily origin destination (OD) networks. We use these networks to study the difference in daily mobility choices arising from residentially segregated areas.This represents the first data driven study to investigate whether racial preferences can be discerned from individual mobility choices. If so, results would indicate that not only do we organize residentially by our socio-economic demographics, but also move according to these as well. With the rapid migration into urban areas, understanding how social decisions influence mobility has consequences in many domains such as epidemiology and urban planning. But perhaps more interestingly, this work also servers as a platform to link individual mobility

data with social theories that posit racial isolation and concentrated disadvantage heighten exposure to criminality and reduce access to resources and opportunities [40, 15]. In addition the ability to know where different races interact can help pin point the definition of "local context" which is used to measure social isolation and racial exposure in sociological works as well as identifying areas of interaction for epidemiologists [37, 26, 48, 49].

1.2 Literature Review: Sociological Studies of Segregation

The study of residential segregation traces its origins to 1926, when Robert Park formally defined residential segregation as the link that exists between both the social distance and physical distance [34], since then both its causes and consequences have been studied extensively both empirically and theoretically. There is a substantial literature concerning the measures of residential segregation and more recently the literature's emphasis has shifted toward the causes and consequences of residential segregation. This is in part because of the application of economic theory to social theory and social dynamics, which have allowed individual behavior to inform models of segregation. We will briefly discuss metrics, and then we will review the literature concerning causes and consequences.

1.2.1 Metrics

The formal definition of residential segregation remained somewhat fluid until 1988, when it was presented as a multidimensional phenomena varying along the distinct axes of evenness,exposure, concentration, centralization and clustering [29], this definition has been widely adopted in the succeeding literature. Each dimension is meant to capture a different aspect of segregation, they are as follows:

- **evenness**: refers to how uniformly distributed each race is over local contexts. If the racial distribution of each local context is the same as the global racial

distribution, evenness is maximized.

- **exposure**: refers to the extent that members of one racial group are exposed to members of another group in each local context.

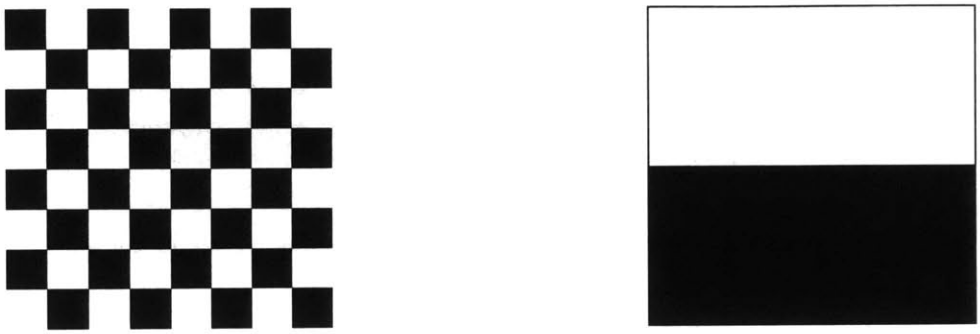
- **concentration**: refers to the amount of relative physical space occupied by a minority group in each local context

- **centralization**: refers to the proximity of the places occupied by minority groups to the city centre.

- **clustering**: the degree of agglomeration of those areas inhabited by a minority group. Concentration refers to one local context, and clustering measures if segregated local context are spatial contiguous.

Recently a discussion concerning spatial vs spatial metrics has brought the distinctness of these dimensions into question. In residential segregation literature, aspatial measures of segregation define a local context for each individual and quantify the extent of how these local contexts differ across individuals. Aspatial measures do not consider the patterning of individuals' local contexts distributed in space however, metrics that do consider the spatial distribution of of local contexts are called spatial metrics. To highlight the difference consider Figure 1-1. On the left each block of the check board represents an all white or all black neighborhood, on the right we have the same number of black and white neighborhoods, but all the white neighborhoods are now on one side of the board and all the black neighborhoods are on the other. An aspatial measure would consider these two cases the same, since the measurement is only at the neighborhood level and the placement of these neighborhoods do not matter.

In [37] it is argued that the distinction between evenness and clustering is simplify an artifact of aspatial measurements at a single geographic scale. They demonstrate evenness at one level, say census tracts, is strongly related to clustering at lower levels, such as block groups. This dependance makes the distinction between the two dimensions arbitrary. They claim the existence of only two dimensions of spatial segregation:



(a) (b)

Figure 1-1: Two different spatial configurations of all back and all white neighborhoods which aspatial measures cannot distinguish between

spatial exposure and spatial evenness and these distinct patterns necessitates the need for two different types of metrics to measure the extent of each axis in1-2.

Spatial incidences not only allow for the quantification of the exposure and evenness they also capture the change in segregation at different geographic scales. Residential segregation metrics are particularly sensitive to what scale is being used, which echoes the criticism that evenness and clustering have an arbitrary distinction. For instance, larger cities are often said to be more segregated than smaller ones but this is based on spurious correlation amongst segregation and city size. Measures based upon census tracts data will tend to report higher values for bigger cities because larger cities have high population densities so their smaller census tracts might cover only one neighborhood, where as smaller cities might have larger census tract that cover several neighborhoods. This bias would be reduced at smaller levels of spatial aggregation. Consequently, it is important to be careful when comparing the level of residential segregation over larger spatial areas. Moreover, it is advisable to carry out a multi-scale analysis. In [36] a methodology to asses the effect of scale on segregation is developed

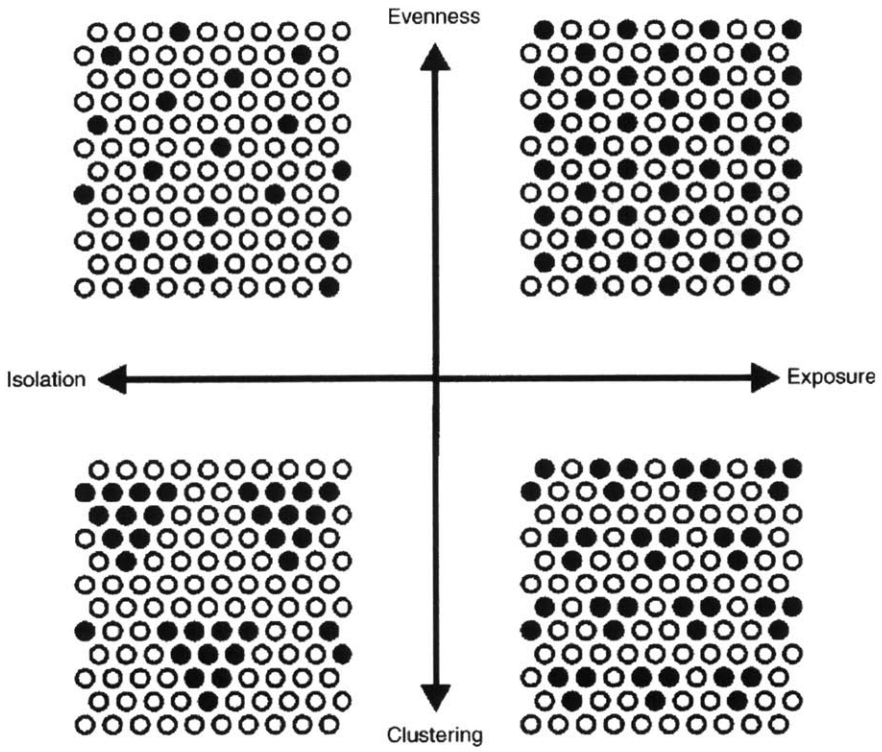


Figure 1-2: Taken from [37] In the upper half are two patterns that have low levels of spatial clustering. In the bottom half of the figure, both patterns show greater clustering then the corresponding patterns above, but roughly the same exposure.

by computing the H-index at various geographic scales, this is called a segregation profile. The H-index is defined as:

$$H = 1 - \frac{1}{TE} \sum_{p \in R} \tau_p E_p$$

Where, E is the entropy of the racial composition of the entire study area, E_p is the entropy of the racial composition in location p, R is the set of all locations considered, T is the total population, and finally τ_p is the population density at location p. The information theory index, H, is a measure of how much less diverse the local environments of individuals are when compared to the total region R. H is equal to 1 if there is maximum segregation, meaning each location is mono-racial, and equal to 0 if each location has the same racial composition as the total area, meaning

that each race is distributed uniformly or complete integration.

1.2.2 Segregation Causes

Segregation is typically presented as a result of endogenous forces such as individual preferences or exogenous forces that sort people across urban areas based on policy and real estate market dynamics. Models of endogenous forces, often cite income as a driver of segregation [32]. These models assume that higher income individuals will have a higher preference for land, meaning that low income households will end up living closer to the central business district (CBD) of a city, this leads to higher concentrations of low income households within the city. This urban structure is subject to change if there are entertainment activities that generate trips to the CBD, such as shopping or young high income individuals prefer to live closer to the CBD [8]. In fact the urban structure will be fully reversed if the demand for accessibility is greater than the demand for space. Another cited endogenous cause of segregation is the willingness to live among peers, or individual preference. The most widely known model of individual preference which leads to segregation is Schelling's model.This model is initialized by randomly placing black and white pieces on checker board. These pieces are then moved to vacant spots based on a tolerance parameter. This process depends on a rule that is previously determined which specifies individuals' willingness to live amongst another race. It is shown even if one group has a slight preference to live near their own race a segregated equilibrium is reached.

Models of exogenous forces cite public policy as a driver for residential segregation. Policies that intentionally force segregation are known as de jure segregation, an example of this is Nazi Germany's policies forcing Jewish ghettoes. Polices that have the unintentional side of effect of segregation are polices that reinforce or encourage the use of exclusionary powers on the part of predominately white and change the nature of the information available to different types of people making location decisions [50]. The second class of public polices that can lead to residential segregation concern zoning and public housing [6, 30], often zoning is used as an exclusion device and public housing which is affordable for lower income individuals are located in clusters

apart from the CDB.

Another model of exogenous forces cite real estate market dynamics as a cause of residential segregation, for example let us consider the process of gentrification. As gentrification increases housing prices and the cost of living increases and lower income individuals are forced to move out, meaning more higher income people move in. These causes of residential segregation only account for a few of the proposed models in the literature, but it is important to note that there is no consensus on what causes segregation and it is likely to combination of all of the above and more.

1.2.3 Segregation Outcomes

In general, residential segregation outcomes have been studied with respect to how it affects segregated populations. Segregation has been citied as reducing exposure to good schools and academic peers, heightening exposure to criminality and creating an environment where segregated individuals do not have access to economic opportunities which perpetuates income disparity[11, 13, 22, 39, 40]. This study concerns itself with another outcome of segregation, which is how it effects the types of locations people visit in their average day.

1.3 Literature Review: Urban Mobility

Studies on human mobility have arisen from various fields of study such as transportation engineering, urban planning and epidemic molding. Here we will focus on statistical physics approaches to modeling individual mobility patterns. In mobile phone studies trajectories are analyzed and statistical measures such as radiation of gyration and jump length distributions are used to characterize how individuals move during their daily life. Studies incorporating call record details are subject to the challenges of data cleaning and storage but provide an unprecedented granularity to study individual trajectories.

The value of mobility reaches far beyond being able to predict geographic movement but provides a whole new framework to consider human interactions from a spatial,

temporal, and contextual perspective [4]. Knowing how people people from one point to another in city provides a new platform to study the differences in daily lives of individuals of different race or ethnicities. Before characterizing behaviors and characteristics that differ it is important to note that human mobility is deeply entangled with a city's geography and spatial economics such as the spatial distribution of points of interests. Thus mobility patterns from cites are not all the same, but we can use models that incorporate these factors such as distance and population to account for these differences. The importance of scale also applies to studies of human mobility, which occurs from short-distance daily travels to long distance air travels. Scale is often incorporated into models of mobility by including distance or a proxy for distance.

1.3.1 Statistical Properties of Human Mobility

In [7] the circulation of bank notes is studied as a proxy for human mobility, where it is show that distribution of traveling distance follows a power law:

$$P(\Delta r) = (\Delta r)^{(1+\beta)}$$

where Δr represents the spatial distance between temporally consecutive locations. They observe that most individuals only travel over short distances. They secondly show that the probability of small spatially confined region diminishes super diffusive spread. In [19] Gonzalez *et al.* analyzed the trajectories of 100,000 anonymized mobile phone users whose position was tracked for a six-month period to show individual mobility patterns differ from the movement of banks notes which can be modeled as Levy flights predicted by random walk models. They demonstrate individual mobility patterns have an exponential cutoff and is better approximated with a truncated power-law:

$$P(\Delta r) = (\Delta r + r_0)^{(-\beta)} exp(\frac{\Delta r}{k})$$

With $\beta = 1.75 \pm .15$ and $r_0 = 1.5$ km and k between 80km and 400km. They further show human trajectories display high degrees of both temporal and spatial regularity and each individual can be characterized by a signature travel distance and signification probability of visiting the same few locations. Radius of gyration,r_g, characterizes individual's average trip length by calculating how far their average trajectory is from the center of the mass defined by all their trajectories, r_g is also approximated by a truncated power law:

$$P(\Delta r_g) = (\Delta r - g + r_g^0)^{(\beta_r)} exp(\frac{r_g}{k})$$

Where $\beta_r = 1.65 \pm .15$ and $r_g^0 = 5.8$ km and $k = 350$ km. This result demonstrates that most people travel in a close vicinity to their center of mass.

Song *et al.* [44] show that exploration and preferential return can be used to recover individual mobility patterns. They show that individuals explore, or visit new locations, with probability $P_{new} = \rho S^{-\gamma}$ and they return to a previous location with the complimentary probability $P_{return} = 1 - \rho S^{-\gamma}$. For specific locations the probability that an individual returns is proportional to the times they have visited in the past. Their findings reveal the following highly non-random properties of individual movement. The number of distinct locations a user visits at time t denoted $N(t)$ is expected to follow $N(t) \propto t^{\mu}$ where $\mu = 0.6 \pm .02$ and the visitation frequency f of the k^{th} most visited location follows a power law $f_k \propto k^{\eta}$ where $\eta = 1.2 \pm .1$

A priori we would assume if residential segregation is highly clustered. individuals would end up visiting locations with similar demographic features simply given their proximity and nature of individual mobility patterns characterized above.

1.3.2 Mobility Studies

There are many studies that link individual mobility patterns to different aspects of social and city structure. In [43] it is shown by incorporating the heterogeneity in population distributions within cities and larger scales mobility can be modeled in a closed analytical form. There are studies linking mobility patterns to the structure of cities [38, 28] and quantifying interaction probabilities within cities [45]. In the social realm [17] mobility patterns of different groups are analyzed to reveal in China women and children travel shorter distances than men. Furthermore in [5] individual visitation patterns are shown to be closely related to the visitation patterns of friends.

Recently network science has been used to investigate properties of locations using origin destination networks constructed from human trajectories. These studies are powerful in revealing dynamic properties of locations while only studying static topological network measures. Node degree and connectedness are two such measures that can provided context on locations. Community detection techniques have leveraged network properties to reveal interesting characteristics about locations and segmentations of the geography that are not visible in the political boundaries[46, 35].

Lastly we outline the only study we found using CDR data to investigate segregation in the different activity spaces of residence, work, and other [48]. Using cellphone data to identify user locations of home and work and preferred language, the authors are able to assign ethnicity to user's as Estonian or Russian. They then go on to measure co-presence of these two ethnic groups at different times and locations. Their analysis indicate that at home and at work, the cellphone users experience varying degrees of segregation but outside of home and work regions the segregation virtually vanishes.

1.4 Thesis Outline

The rest of the thesis will create an experimental framework to conceptualize and quantify the effects of segregation on human mobility patterns.

In Chapter 2 and Chapter 3 we develop a methodology to incorporate individual travel patterns into segregation metrics to better capture and characterize daily

environments experienced by individuals. These reformulated metrics allow us to glean interesting information about the importance of different locations to different races and quantify the effect of trip distance on the exposure rates for different races.

In Chapter 4, we utilize two rank based models that to explicitly test the effects of residential segregation on mobility patterns. In one model we use as input total population, in the other we use only specific race populations to predict geographical movement. We find the race blind model, which uses total population, reproduces the empirical data quite well and the race aware model that uses only one race's population distributions to predict that respective race's movement improves performance slightly. These simple models provide a very powerful platform to understand the impact that geographic patterning of segregation on the daily lives of user's who live in these segregated areas.

The final chapter concludes this thesis.

Chapter 2

Data Description and Experimental Setup

2.1 Introduction

The pervasive use of cell phones has made CDR data an effective sensor of individual mobility at an unprecedentedly granularity temporally and spatially. This data is generated continuously and unobtrusively, as natural output of our daily life but unlike data generated in labor intensive ways, like travel surveys, this data is not neatly packaged, rather it is messy, incomplete, and enormous. Thus, it requires careful preprocessing and treatment to extract meaningful information and leverage its potential. In [14] population densities are estimated spatially at multiple time scales using mobile phone traces, demonstrating the efficacy of CDR data in mapping population level dynamics. Furthermore in the transportation realm, methodologies have been developed to treat CDR data to create OD trips that match Travel Surveys [1, 47]. These methodologies and validations demonstrate the use of CDR data to efficiently mine the travel patterns of entire populations. In this chapter we outline the CDR treatment and upscaling process by which this is possible. Since both OD accuracy and segregation are sensitive to scale, we also evaluate a proper scale to conduct our study and finally we choose two metrics, entropy and exposure, to measure segregation.

2.2 Data Description

2.2.1 Census Data

We use population tables from the 2010 American Community Survey (ACS) which can be accessed through the US Census. The ACS randomly samples addresses in every state in the US, around 1 in 38 households receive an invitation to participate in the survey itself. After the data is collected, demographics are scaled up to represent the entire population. Demographics are provided at many spatial scales, but census tracts represent the smallest territorial unit for which population data is available. The U.S. Census Bureau designs the tracts to be homogenous with respect to population characteristics such as economics status, living conditions, and population [9]. This is done as a means of creating statistically comparable geographic units. The census tracts in our study region have an average of 4,809 inhabitants.

Below we examine racial and population characteristics of the 974 census tracts in our study region. In Table 1 we calculate how many census tracts are predominately of the races we consider: white, black, hispanic, and asian. We plot the racial distributions within the census tracts and give the total populations in the legend of 2-1. Whites are by far the majority in population and there are very few census tracts with a majority of minorities.

In figures 2-2 and 2-3 every dot represents 100 people of a particular race. For each census tract, dots are generated and labeled with races according to their respective racial population distribution described above. For the visualization, dots are uniformly distributed within the borders of corresponding census tracts and colored by the race they represent. We are left with a visual approximate of the racial composition of our study area and a zoom in of the Boston Metropolitan area. These maps show our study region is predominately white and fairly segregation. For quantification of this segregation see Section 2.4.

2.2.2 Cell Phone Data

Our CDR data set contains more than 8 billion anonymized mobile phone records, obtained from several phone providers, covering approximately 2 million users in our study area. The data spans two months in the spring of 2010. Despite the two month coverage of the data set, only a few users are observed during the whole 60 day period due to reindexing by the provider on the 17th day, other users are observed for at most 17 or 43 days.

For each record we are given the following information: an anonymized user identification number, the latitude and longitude of the record, and finally the timestamp at the moment of phone activity, this activity encompasses calls, text messing, and web browsing. Typical cell record datasets are given with respect to cell towers, in our case, the provider estimates the location of reach record using a triangulation scheme, resulting in location accuracy of 200-300 meters.

2.2.3 Massachusetts Household Travel Survey

The Massachusetts Household Travel Survey (MHTS) contains information on 153,099 trips made by 32,739 people from 15,000 households [33] during June 2010 and

MA Census	# tracts
Total	974
Dominantly White	884
Dominantly Black	52
Dominantly Hispanic	24
Dominantly Asian	6
No population	8

Table 2.1: Census Tract Race Characteristics

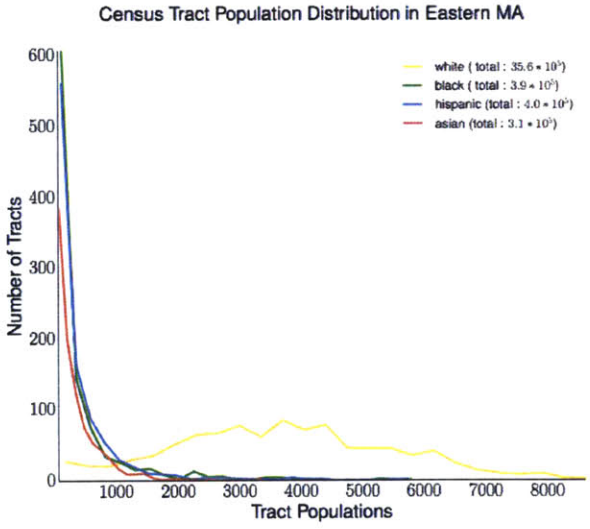


Figure 2-1: Population Distributions

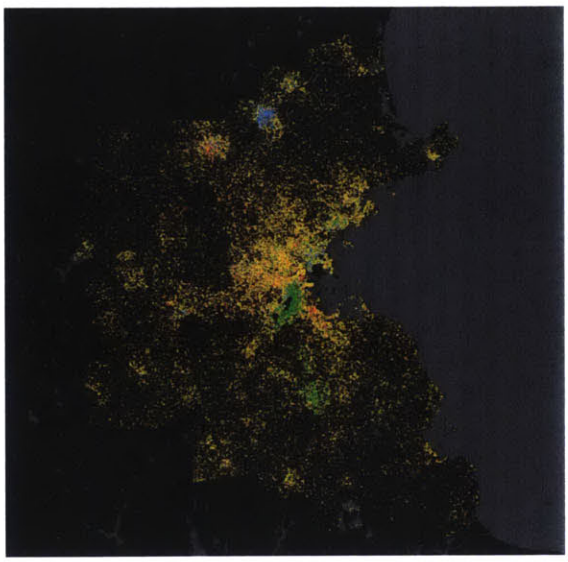


Figure 2-2: Study Area Distribution of Racial Groups.Yellow represents white, Green represents black, Blue represents hispanic, and finally Red represents asian.

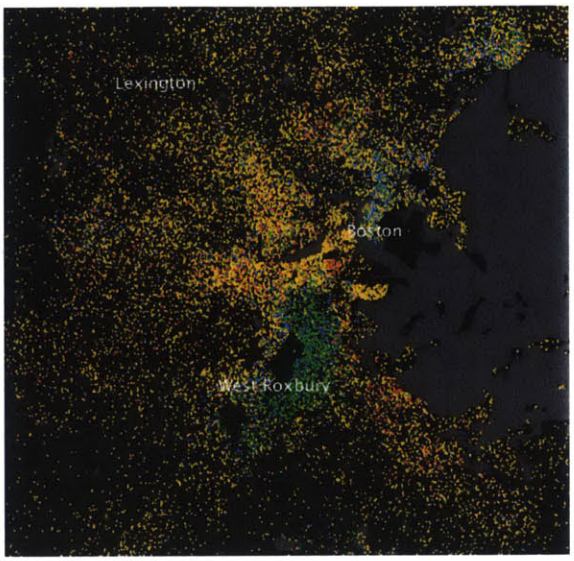


Figure 2-3: Boston Area Distribution of Racial Groups. Same color scheme applies

November 2011. Each member of the household prepared a diary for one specified day, they were asked to report all trips, models of travel, prices paid, and types of activity at each visited location from the beginning to end of the day. The survey also provides a scaling factor for each trip which we use to expand the survey data to represent population level travel.

2.3 CDR Data Processing

2.3.1 Stay Extraction

The first step in the data mining process is to identify 'stay locations' which are the locations where users engage in some activity. These stay locations can be distinguished from pass by locations which are records made while traveling. We use a variant of the

Hariharan and Toyama method [23] detailed in [25]. Once a stay point is identified its location is set as the centroid of all the records belonging to that stay. The next step is to identify stay regions from stay points. This is done because many different stay points, identified from the same user's different trajectories, may in fact refer to the same location. This step is done using a grid based clustering algorithm. In this study, the maximum stay region is set to 300m to approximate the location accuracy. Finally, once stay locations are determined we impose a duration criterion and consider only those stay locations where we know the user has been for at least 10 minutes.

2.3.2 Activity Inference

Since trips are induced by purpose, it is necessary to fully understand the those purposes before any analysis can be done evaluating personal preference. Some trips are done out of necessity, like dropping your children off at school or going to work, where as other trips like those for leisure activity are done from personal motivations. Using the same methodology developed in [1] we assign labels to the stay regions described above in order to generate trips by purpose. The stay process yields a timestamp and duration for each stay location. Using this for each user, we are then able to assign an activity type of home, work, or other to each stay location. Home locations are defined as the most visited location on both weekends and weekdays between the hours of 7pm and 8am. Since work hours are typically between 9am - 5pm we expect users to home between 7pm and 8am given a buffer time of travel and after work activities. In essence we make the simple assumption that people spend the night at their own homes.

For each user, work is defined to be the location, $l \in L$ that satisfies (2.1).

$$l \equiv \arg\max_{l \in L} d_l * n_l \tag{2.1}$$

where:

L : the set of all stay locations visited by the user in consideration

d_l : great circle distance of location l from the user's home

n_l : number of times a user has been seen at location l on weekdays between 8 am and 7 pm

This definition of work is matches the pattern that work trips are likely to span longer distances than non-work trips found in the MHTS.Once work has been established for each user the following filters are applied. If a user visits work less than 8 times, or the distance from the user's home to work is less than 0.5 km, the stay region is switched from 'work' to 'other'. As a result, not all users in our dataset are given work locations. These filters prevent the false identification of work based on infrequency of visitation (which does not allow for certainty in classification) and or proximity to home location (which could simply be noise in the signal).

2.3.3 Filtering and Expansion

To capture only those users who's home location is adequately represented in the CDR data, we filter out any user who we have captured less than 8 times in their home stay location. Accurate home location assignment is absolutely necessary in following upscaling methodology. This filtering results in 335,795 users: an order of magnitude larger than in most household travel surveys.

Next we upscale user trips to represent population level travel patterns we locate the census tract containing each user's home location. Once this has been done for every user, we have calculate an expansion factor as the ratio of residents as identified in the CDR data and the census population from the ACS. If any tract has fewer than 10 CDR residents, the scaling factor is set to 0, so prevent the use of tracts where the CDR data is under-representative.

2.3.4 Inferring Departure Times

The stay process yields a timestamp and duration for each trip but this is only an observation based on phone usage, that does not accurately correspond to when the trip was initiated. Rather than using these times as departure times, we infer trip departure by creating probability distributions from the 2009 National Household Travel Survey (NHTS) and sample this distribution to account for this uncertainty. We generate six hourly distributions for weekdays and weekends for the following trip segmentations found in transportation literature: home-based-work (HBW), home-based-other (HBO), and non-home-based (NHB). For each user, we sample the corresponding distribution within an interval corresponding to two of their consecutive stays occurring within a 24-hour period. Furthermore, if we do not observe a user's last trip of the day to be at home, we create a new trip so they return home. Through this process, we construct trip departure times on all days we observe each user.

2.3.5 Validation

We validate our departure process by comparing the CDR departure times with the MHTS departure times to see that they match 2-4. We favorably find that the trip departures seem to follow similar patterns with the noted exception of consistently more CDR trips in the late night hours and smoother distributions as a result of the sampling. While the abundance of trips in the evening compared to the survey may be due to a slight mismatch between the frequency of calling and trip-making throughout the day, it may also highlight an advantage of CDR data to capture late night trips not typically reported in survey responses of an average day.

In 2.2 we compare the relative share of trips in each trip type category. Regardless of the more trips being reported in the MHTS we find when we aggregated to the town level we find almost perfect correlation between the two data sets. This suggests that our inferences of home, work, and other activities and upscaling methodology seem reasonable. We will further examine the effect of aggregation on correlation in a subsequent section.

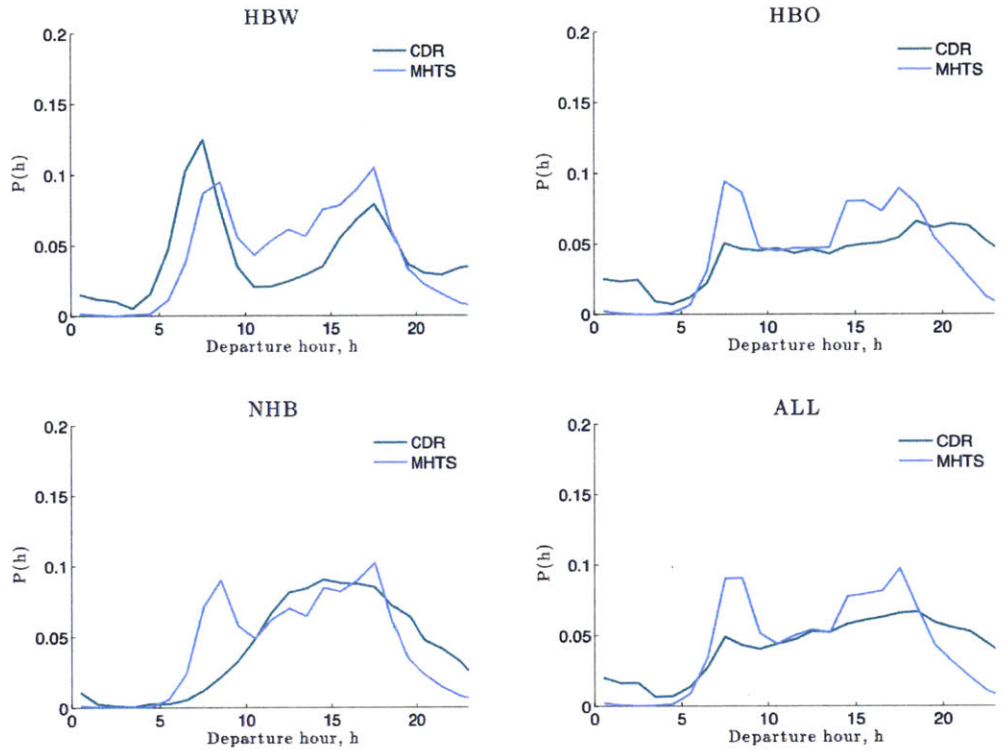


Figure 2-4: distributions of hourly departure times for HBW, HBO, NHB, and total average weekday trips in the CDR and MHTS datasets

	HBW	HBO	NHB	ALL
CDR (x 10^6)	2.8	7.8	4.7	15.3
MHTS (x 10^6)	2.0	8.5	6.7	17.2
Tract correlation	.26	.64	.59	.59
Town correlation	.96	.97	.98	.98

Table 2.2: Pearson correlations between MHTS and upscaled CDR for trips by different purposes and aggregation levels

2.3.6 User Race Assignment

From the above we can be fairly sure that our assignment of user's home tract is fairly accurate so we assign races to our users based on the racial distribution of their home tract. We consider two options, first probabilistic assignment: given the racial distribution of a given tract as [%white, %black, %hispanic, %asian], or for illustrative purposes lets assume the distribution of tract A is [0.7, 0.2., 0.1, 0.0] we can assume

each user living in that tract is 70% white, 20% black etc. When we consider their trips, the same trip will have some probability of being each certain race. Although this is a valid option to asses the expected value of all trips that minorities could make, we opt to simply assign the majority race of each tract as our user's race. So in the case of tract A, all users would be white.

In our study we want to asses the effects of segregation, more important to us than assigning a user's race accurately, is observing the trip behavior emerging from residentially segregated areas. If we choose to assign race probabilistically it would dampen the effect of segregation we seek to measure. Each method of assigning race has its downfalls in the obvious misclassification of user race. Thus the race assignment in this study should not be thought of as the user's actual race, but as a user who lives in a neighborhood dominated by their given race.

2.4 Segregation and Scale

Here we address the effect of scale on segregation in our study area as well as the accuracy of our ODs. As noted before, aspatial segregation metrics rest on a problematic assumption that the two cases in 1-1 are the same, they do not take into consideration proximity and are subject to change given different scales. Much segregation literature relies on the aspatial metrics that use census tracts as their local contexts. The popularity of census tract arises from is availability and their construction with respect to the statistical features mentioned above. Furthermore census tracts are thought by many to represent real neighborhoods. Although the tracts are drawn with statistical considerations in mind, the borders often follow political boundaries such as town, county, and state lines. These boundaries were drawn and stabilized decades ago, and have remained relatively static since, making some census tracts seem arbitrary. Conventional census tracts studies face three major criticisms [26]:

- census tract based metrics assume that each tract constitutes an appropriate unit for capturing segregation, but does not account for the potential variation

among regions. census tracts cannot distinguish when segregation changes over short distances or larger distances.

- variation exists in the geographic area of tracts, which means the 'scale' of the tract-based segregation is ambiguous

- treating tracts as spatially discrete has the consequence of treating all persons within the same tracts are proximal but completely disjoint to residents outside the tract, even at tract boarders

In [26] present the a methodology to overcome these issues. They superimpose a gridded mesh onto the region and compute the racial distribution of the cells as a weighted average of the population distributions from the tracts which intersect the cell in consideration. Then they use a two-dimension biweight kernel function that incorporates distance decay, thereby assigning more weight to closer cells, to create a local context which can be used in any spatial segregation metric. This is an elegant solution but cannot be applied in our case. Since we are working with origin destination trips we need our level of aggregation to partition the space into discrete units. Secondly we address the concern of varying geographic areas in census tracts. The authors argue that the scale at which segregation is measured is ambiguous when using census tracts since regions less densely populated are spanned by larger census tracts. In our study region, we do see 2-5 there are a few census tracts that have significantly larger area then the rest.

If we consider, however, what segregation indices are trying to gauge, which is an individual's context it seems relevant to see the average travel distance for each census tract. Below in 2-6 we see people who live in less densely populated areas make much longer trips than people who live in densely populated areas. This would mean their local context does in fact span a larger geographic area. Thus it might be better to use varying geographic areas based on population density to capture local contexts, and census tracts do just this, perhaps even better would be incorporating varying levels of distance decay into the kernel functions. Nevertheless, we choose to use tracts

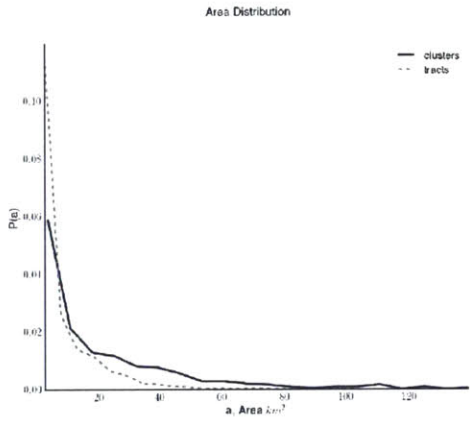


Figure 2-5: There are few tracts and cluster with larger areas

in this study for their simplicity, but the metrics we use are easily extended to using their spatial forms, which could incorporate the methodology proposed in [26] .

Figure 2-6: Trip distance decays exponentially as population density increases in Home-Other trips

Although census tracts as our basic unit we examine the effect of scale by moving to different levels of aggregation using k-means clustering. K-means clustering begins with the choice of the number of clusters n. For various choices of n, we randomly pick n (lat,lon) coordinate pairs in the study region to represent the centers of the clusters. They are denoted as k,$k = 1, ..., n$. Each tracts's center location is denoted as a vector X_i, for $i = 1, ..., 974$. The goal is to find an assignment of X_i to clusters, as well as a set of vectors μ_k, such that the sum of the squares of the distances of each data point X_i to its closest vector μ_k, reaches a minimum. We use a 1-of-K coding scheme to represent the cluster where data point X_i currently belongs. For each data point X_i, we introduce a corresponding set of binary indicator variables $r_{ik} \in 0, 1, k = 1, ..., n$ indicating the cluster location of data point X_i. Namely, if data point X_i is assigned to cluster k then $r_{ik} = 1$, and $r_{ij} = 0$ for $j \neq k$. We minimize the objective function, J, which is the squares of the distances of each data point to its assigned vector μ_k, where J is written as:

$$J = \sum_{i=1}^{974} \sum_{k=1}^{n} n||X_i - \mu_k||^2 \tag{2.2}$$

The distance $||X_i - \mu_k||^2$ is the haversine distance between the two coordinate pairs. To update r_{ik} and μ_k iteratively we perform the following two steps until convergence.

1. While keeping μ_k fixed find the values of r_{ik} which minimize J. That is, finding the closest cluster to each data point.

2. While keeping r_{ik} fixed find the values of μ_k that minimize J. Since J is a quadratic function of μ_k we take the derivative of J with respect to μ_k and set it to zero and solve for μ_k.

We must note that when using k-mean clustering, the resulting clusters will also have varying areas 2-5 since the tracts they are comprised of have varying areas. Thus the clusters near the center of Boston are much smaller than the clusters in the western part of our study region. We believe this area distribution is more reflective of

individual's local context than uniform areas as argued above. Once we have k-means clusters at serval values, we examine how the segregation profile and OD correlation is effected by scale. As we aggregated to form larger but less clusters, the accuracy of our CDR OD trips increase, but the segregation profile decreases. This makes sense because we are losing very small range trips which could be noise or which people may not necessarily report in a survey. But as we aggregate we are in essence taking the already small minority population in Boston and uniformly distributing them within their respective clusters, making Boston seem more mixed than it is, so the H-index decreases 2-7. This of course is the nature of aggregation and segregation will be measured as concentrated on lower scales.

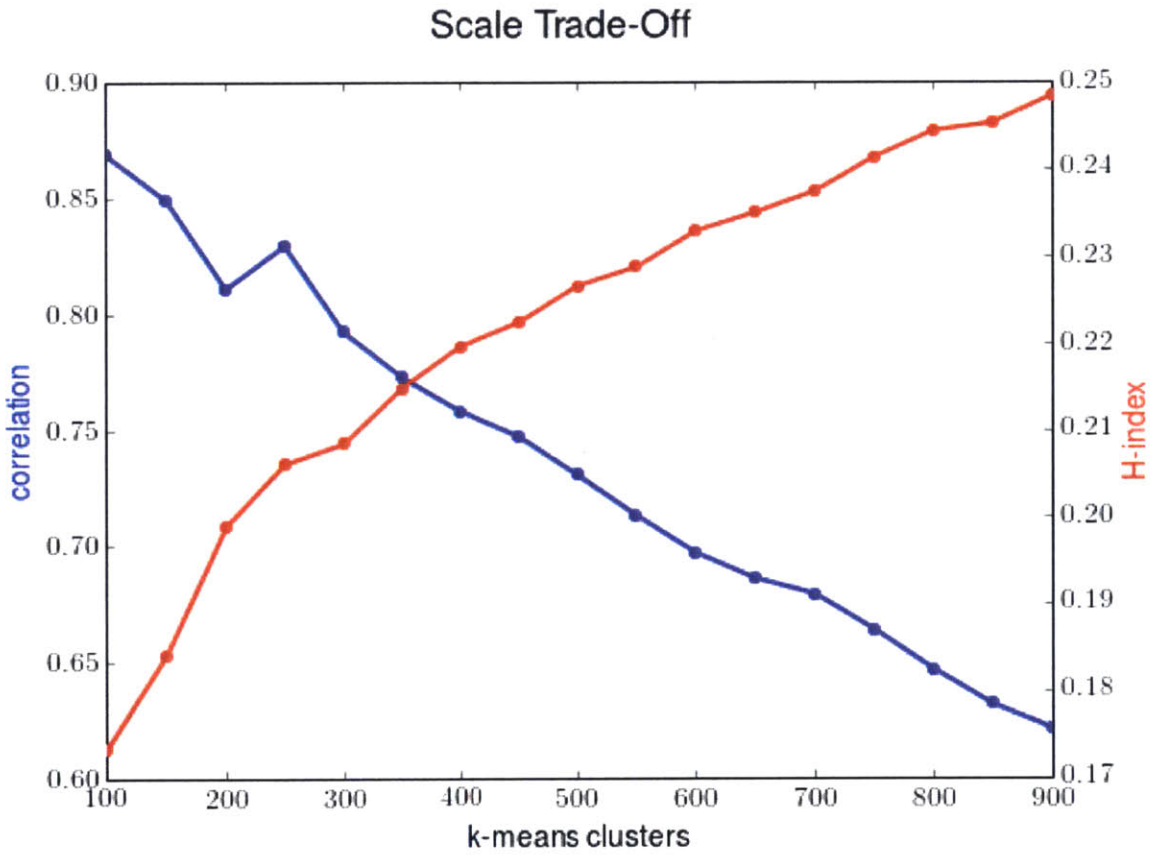


Figure 2-7: The problem of scale in our study, as correlation increases our measures of segregation decrease

We choose to carry out our study on the level of the 350 k-means clusters to gain some accuracy in our ODs but not allow the H-index to decrease substantially. We

demonstrate how entropy and exposure change from the tract level to 350 cluster level in the next section to provide a comprehensive evaluation of scale choice.

2.4.1 Metrics

Since we are not interested in one aggregate value of segregation but instead characterizing the locations in individual trajectories we consider the individual contribution of each tract and cluster. We use the values of entropy and exposure to characterize each cluster. We compute the racial distribution of each cluster as the weighted average of the racial distribution of each tract that cluster contains. Entropy is defined has the individual contribution of each location to the H-index described in the segregation literature review. Given the racial distribution of each location, which gives the proportion, π_{ir}, of each race r within location i, then the entropy of location i is written as follows:

$$E_i = \sum_{r \in R} \pi_{ir} \log_4 \pi_{ir} \tag{2.3}$$

We adopt the methodology in [24] to jointly understand segregation and diversity through the examination of location entropy. The authors present a methodology to characterize a racial continuum enabling a reinterpretation of the racial landscapes of metropolitan places. They argue that segregation and diversity must be jointly understood and not treated as binary opposites. We adopt their framework and provide maps below to illustrate their characterizations of our study area in figures 2-8 and 2-9. Locations are characterized by "low diversity", "moderate diversity", and "high diversity" based on their entropy scores.

The entropy limits of these labels are calculated based on bounding the racial composition of each location. In our study, where we consider four racial groups, "low diversity" locations have entropy value $\leq .42$. This value is determined based on two criteria. First, this is the maximum entropy that can be reached if one of the four racial groups constitutes 85% of the population- i.e. each of the other 3 groups

constitutes exactly 5%. Secondly, if no group exceeds 80% a value less than .42 can be reached. For example if a location is 75% white and 25% the entropy is 0.4, although no group meets the traditional measure of group dominance of 80%, the tract is not diverse as a whole. Those locations that are labeled "high diversity" have entropy $\geq .76$. This limit ensures that no group constitutes more 45% and the tract's top two groups have a combined percentage of $< 80\%$. Locations labeled as "moderately diverse" are those locations not captures by the other two categories.

In figures 2-8 and 2-9 we present the entropy of tracts and clusters, so we can examine our study area in terms of segregation and diversity, as well as contrast the difference between the two scales.

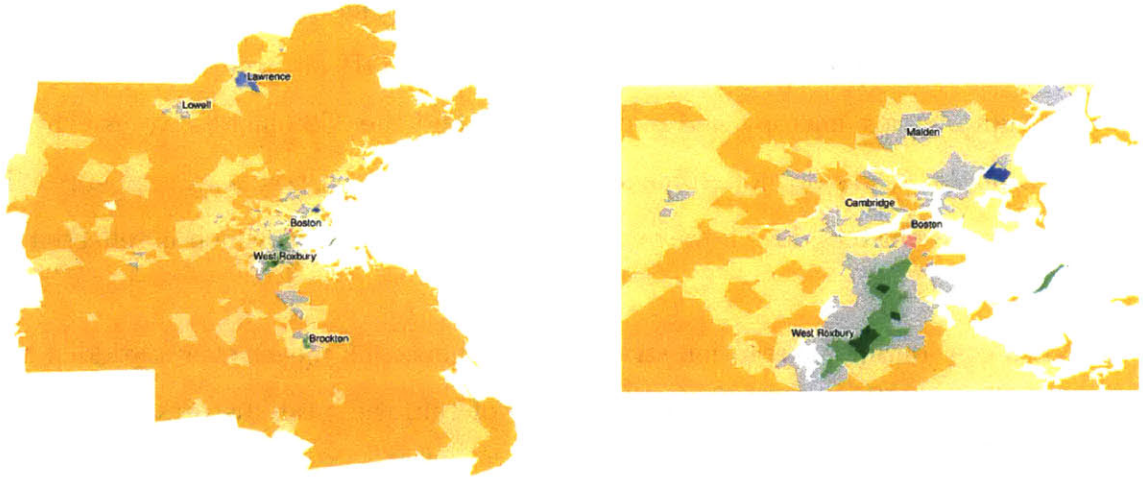


Figure 2-8: Entropy of tracts in our study region. The same color scheme we have used thus far applies, dark yellow refers to "white majority and low diversity", light yellow refers to "white majority" and "moderate diversity" and this convention holds for all races. Grey indicates "high diversity".

Entropy is a measure along the evenness dimension of 1-2, to measure the other axis we use the measure of exposure. Exposure is a measure used to capture the average percentage of race m present in the local environments of race n, denoted ${}_nP_m$. In aggregate the measure of exposure is:

$$ {}_nP_m = \sum_{l \in L} \frac{\tau_{lm}}{T_m} * \pi_{ln} \tag{2.4} $$

Where L is the set of all locations, T_m is the total population of race m in the whole study area, τ_m is the population of race m in in location l, π_{ln} is the proportion of race n in location l. As before, we present the exposure at the tract and cluster level to highlight the effect of aggregation on our metrics of segregation 2.3 .

2.5 Discussion

Here we presented a method of treating CDR data to represent population scale travel patterns. We detail this process because establishing accurate home locations is also used in assigning a race to our users. This race should not be thought of as the user's actual race but more a racial representation of their home tract.

We also carefully consider the effect of scale on OD accuracy on two different segregation metrics, entropy and exposure. We find the 350 clusters we use through the rest of the study do not differ substantially from the tracts in exposure. The aggregation to clusters does however change levels of diversity which is reflected in a lower H-index score.

When considering aggregation strategies the choice of k-means over a gridded mesh scheme was deliberate. Even at the fine scale of 1km the mesh distorts geography as

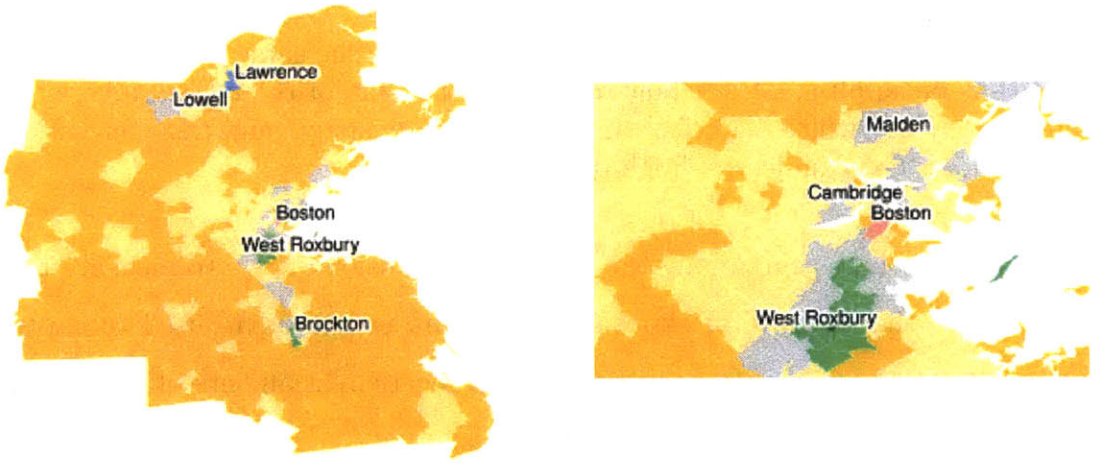


Figure 2-9: Entropy of 350 k-means clusters in our study region. There are no longer any minority majority with "low diversity" clusters because of redistribution of tract populations within the clusters.

race x	race y	tract	cluster
w	w	0.83	0.82
w	b	0.05	0.05
w	h	0.06	0.06
w	a	0.06	0.06
b	w	0.47	0.50
b	b	0.32	0.29
b	h	0.15	0.14
b	a	0.06	0.06
h	w	0.54	0.57
h	b	0.14	0.14
h	h	0.25	0.22
h	a	0.06	0.06
a	w	0.69	0.71
a	b	0.08	0.08
a	h	0.08	0.08
a	a	0.15	0.13

Table 2.3: Exposure of all 16 race combinations at the tract level and cluster level. At the cluster level minorities are more exposed to white and less exposed their own racial group

well as the racial compositions 2-10 and for substantial OD correlation gain we would have to aggregate these cells to coarser levels compounding these problems.

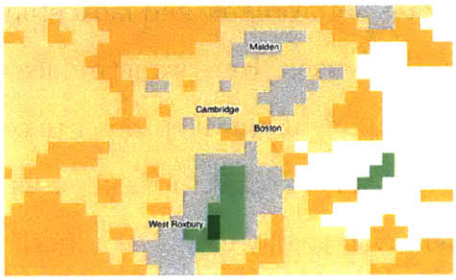


Figure 2-10: gridded mesh, each cell is $1km^2$, imposed over the Boston area and the entropy of the resulting racial distribution

In this chapter we have spent time characterizing local contexts at the tract and cluster level and getting validated OD by purpose. Now we can couple these two and begin to characterize characterize how the visitation patterns emerging from residentially segregated area's differ by examining the local contexts they choose to visit.

Chapter 3

Mobility by Race and Extensions of Segregation Metrics

3.1 Introduction

The goal of segregation metrics is to characterize the local environments of individuals and assess how these environments differ across races. In the perfect study, we would be able to construct ego-centric environments for every individual and would characterize these. This data is not available and most studies have only considered the aggregate data available in the census. Here we introduce data on upscaled individual trajectories into these metrics taking us one step closer to constructing these ego-centric environments of an ideal study. These trajectories allow us to understand the locations that individuals visit in their average day giving us a much more comprehensive idea of what they are really exposed to rather than only characterizing the local environment of their home locations.

We begin by showing that most people's travel patterns display remarkable regularity, with many visits to only a few locations which demonstrates that an individual's trips in an average day will be representative of their experiences in aggregate. Once we have this established, we reformulate the measure of entropy and exposure to incorporate the racial compositions of the locations visited. We argue these reformulated metrics more closely approximate the environments of individuals as well as reveal

interesting characteristics of the locations themselves.

3.2 Mobility Metrics

We begin by measuring a few standard mobility metrics to demonstrate the highly non-random behavior of individual's aggregate mobility patterns 3-1. Firstly we measure each user's regularity, $R(t)$, which is defined as the probability of finding the user at his/her most visited location during hour t. R is a lower bound on the predictability of each user where as tighter bounds would consider temporal correlations. We see that R is time dependent but is relatively stable across users and races. R peaks at night when people are most likely home and reaches its minima in the afternoon. To complement this measure, we calculate the total number of distinct locations $N(t)$ a user has visited each hour and we can see the moments of low regularity correspond to significant increase in N, and similarly N attains it minimum when R reaches its max. These measures show us some temporal characteristics of mobility: that individuals tend to be home at night and travel during the day and are consistent with those in [44]. The bottom two plots $f(k)$ and $P(N_i)$ show us that users spend most of their time in few locations.N_i is the number of locations visited by user i in the complete data set, and $P(N)$ is the probability of finding a user who has visited N locations, we see that most user's have only visited 10 locations during the two month span of our data set. The plot of $f(k)$ measures how often a user visits their k^{th} most frequented locations, it shows us user's frequent only a few locations and rarely go to many others. Since individuals tend to be highly regular we can characterize these few locations they visit to obtain a representative picture of where they spend their time.

3.3 OD Network Generation

Here we construct OD networks to measure aggregate patterns arising from residentially segregated areas. We also use features of these networks as input for our reformulated segregation metrics. We construct the network as follows: each node represents a

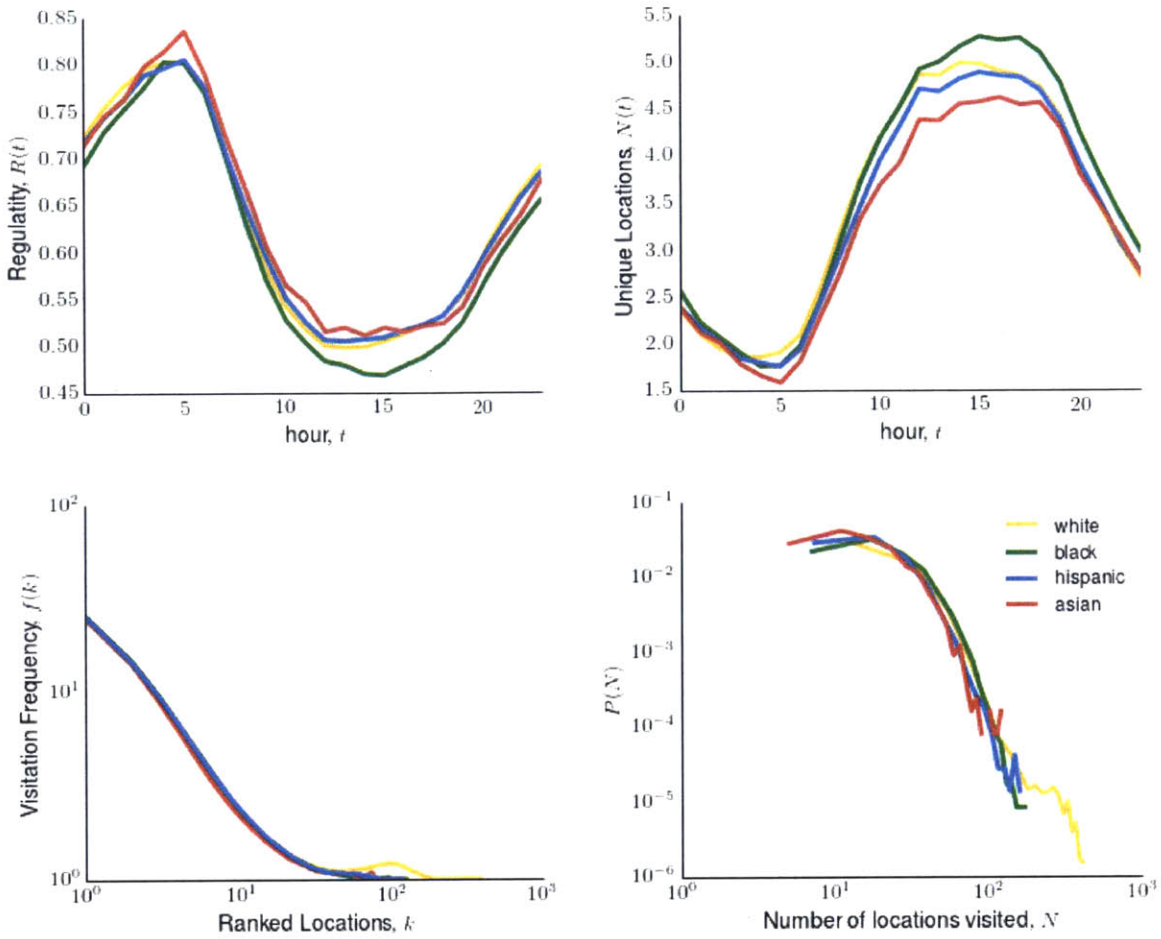


Figure 3-1: Regularity measures: $R(t), N(t), f(k), P(N)$

particular cluster, and edges are present between nodes if a trip has occurred between the two locations. The weights on the edges t_{ij} are the sum of daily flows from $i \to j$, hence the network is directed. The cluster network has 350 nodes, the weights t_{ij}, are found from the trajectories from the same users discussed in Chapter 2. We focus on the average daily HBO trips based on the belief that these trips are most likely to stem from individual choice, but the methodology we develop can be applied to any trip network.

In fig 3-2 we see that the networks arising from the predominately minority areas actually stay pretty close to those areas, and many of the longer range connections are between other area's that have higher densities of the same respective minority. For example we see connections from predominately asian sections of Lowell to Malden

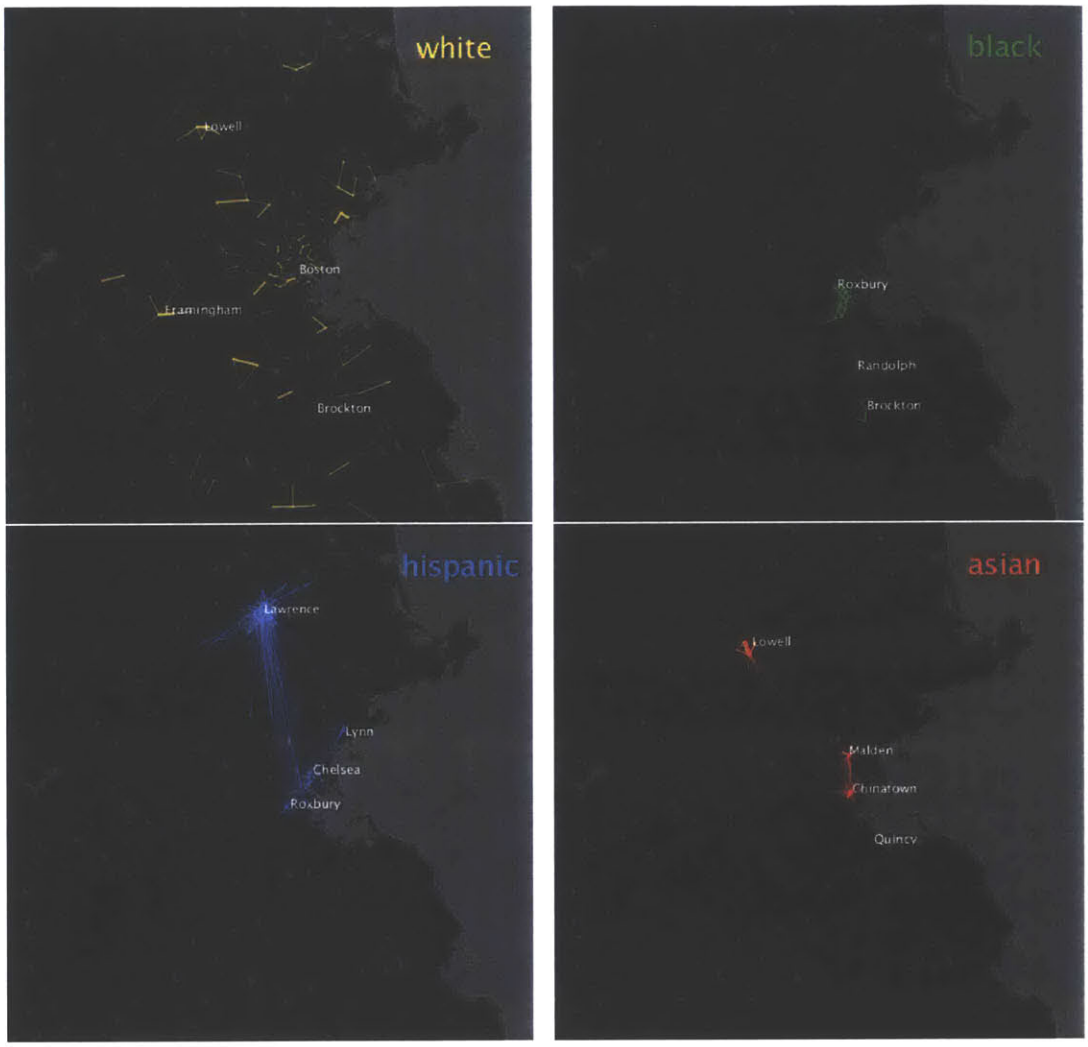


Figure 3-2: HBO OD Networks by Race, the edge opacity and width is determined by the weight on that edge.

where there is another large population of asians.

3.4 Segregation Metrics Reexamined

We reformulate traditional measures of segregation, namely entropy and exposure to include dynamic properties of individual mobility. These measures are often cited as proxies for understanding the local environments of minorities, but a more accurate understanding would incorporate data on where these individuals go throughout the day. Nevertheless, to retain the simplicity of the exposure and entropy metrics, instead

of creating unique metrics for each user and averaging over all individuals, we use topological properties that characterize the dynamics of our OD networks as input, which previously used static measures of population distributions as input.

3.4.1 Entropy

We reformulate the entropy metric to include a more accurate picture of people's whereabouts in order to characterize locations. In the pervious chapter, we jointly considered the diversity of each location as well as geographic segregation by coloring the locations by their majority race.Instead of measuring the static distributions of locations, we measure the diversity of visits to each location formulated as follows:

Let L be the set of all locations. Firstly, for each location $i \in L$: we compute s_i^r, the incoming strength in the OD-network defined by race r. Next, we normalize by total incoming strength in r's network , s_{total}^r, this gives us the overall importance of location i in each race network. Because we have many more white user's than minorities normalization ensures we weight each network equally and are measuring the locations importance in each network, rather than absolute visitation values. Finally we compute the entropy, E_i , of these normalized values.

$$s_i^r = \sum_{j \in L} t_{ij}^r$$

$$E_i = \sum_{r \in R} \frac{s_i^r}{s_{total}^r} \log_4 \frac{s_i^r}{s_{total}^r}$$

Downtown Boston and South Cambridge are shown to be important locations in all race networks but there is not much diversity in visitation elsewhere 3-3. This would indicate that downtown is a place where all the races might encounter one another, in other words, locations with high exposure.

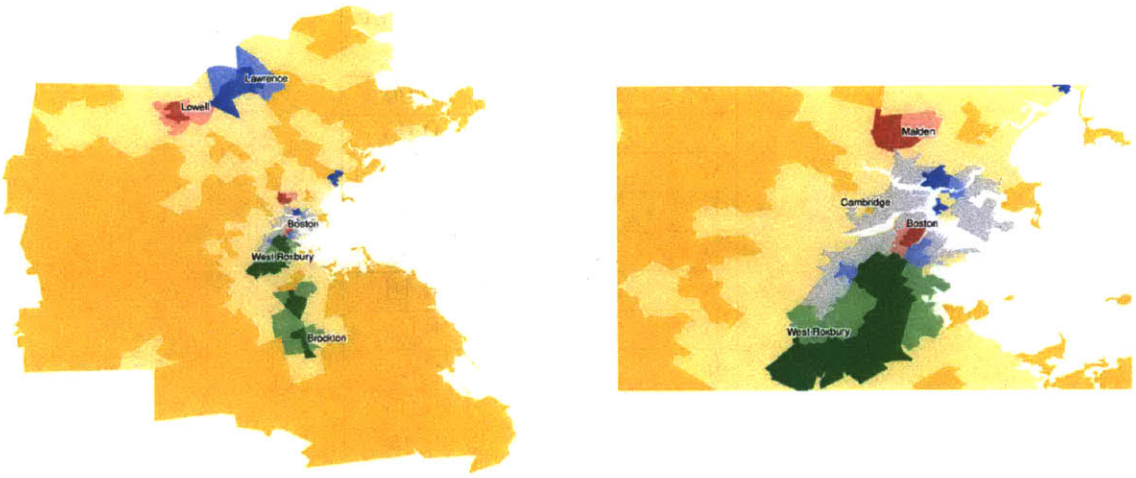


Figure 3-3: Entropy of each location's visitation patterns. Dark yellow refers to a location that is most important in the white network and "low diversity", light yellow refers to a location that is mostly important in the white network but "moderate diversity" in network importance. This convention holds for all races. Grey indicates "high diversity" meaning locations that are important in all networks.

3.4.2 Exposure

Exposure is traditionally a measure used to capture the average percentage of race m present in the local environments of race n. Here we measure a weighted exposure of all destinations in each race network and further examine exposure as a function of distance. Recall that π_{im} is the proportion of race m in location i, we reformulate exposure using the same notation as for entropy:

$$ {}_nP_m = \sum_{i \in L} \frac{s_i^n}{s_{total}^n} * \pi_{im} $$

In 3.1 we provide the exposure of race x to race y measured using the incoming strength of all destinations in the OD network defined by race x.

It is interesting to notice all races are more exposed to white while traveling than in the static measures 2.3, while same race exposure remains significant. Next, we examine how exposure changes given trip distance. Given that most most trips in our OD networks occur between 0 - 5 km with very few above 25km 3-4, we measure

race x	race y	cluster
w	w	0.79
w	b	0.06
w	h	0.06
w	a	0.07
b	w	0.30
b	b	0.47
b	h	0.16
b	a	0.05
h	w	0.41
h	b	0.05
h	h	0.48
h	a	0.03
a	w	0.52
a	b	0.08
a	h	0.10
a	a	0.28

Table 3.1: Exposure of all 16 race combinations averaged over all the destinations in each races' OD network.

exposure in the following intervals $d = [0, 1, 5, 10, 25, \infty]$. For each interval $[d_k, d_{k+1})$ the exposure is formulated below:

$$s_i^r = \sum_{j \in L} t_{ij}^r \qquad \text{if } d_k \leq h(i,j) < d_{k+1}$$

$${}_nP_m = \sum_{i \in L} \frac{s_i^n}{s_{total}^n} * \pi_{im}$$

In figure 3-5 as distance increases exposure to white people increases and exposure to one's own race decreases but remains somewhat high also exposure to other minorities does not increase at any distance. This seems like it might be inflated due to the inclusion of home locations but we see similar exposure patterns in the NHB network 3-6. We do not include all the same analyses for the NHB network for the sake of brevity, but it is worth noting these patterns appear with or without the inclusion of home locations.

Threshold models used in segregation literature often cite that there is tolerance

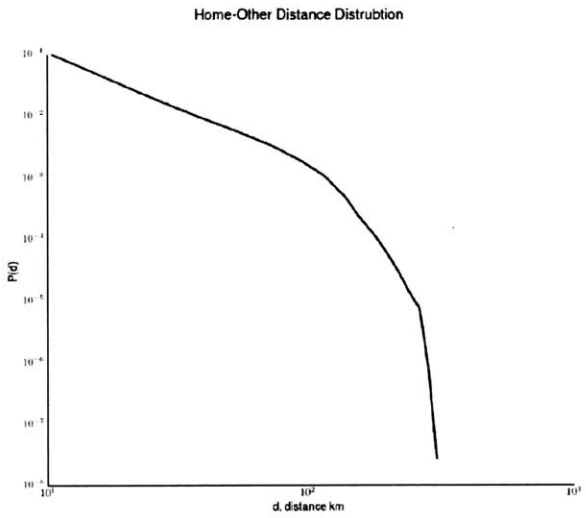


Figure 3-4: Probability distribution of HBO trip lengths

parameter governing residential movement. Inspired by this, we create probability of visitation functions with respect to the racial composition of local environments. What we see in 3-7 seems intuitive, each race has a higher probability of visiting a cluster that has a larger proportion of their own race.

From 3-5 and 3-7 we see a stark pattern emerge: even during daily travel races are likely to visit locations that are dominated by their own race. Although we have looked at exposure by distance we need to fully untangle how much of these effects are due to the geographical clustering of residential segregation. We address this in following chapter using two different mobility models that implicitly incorporate geography and population.

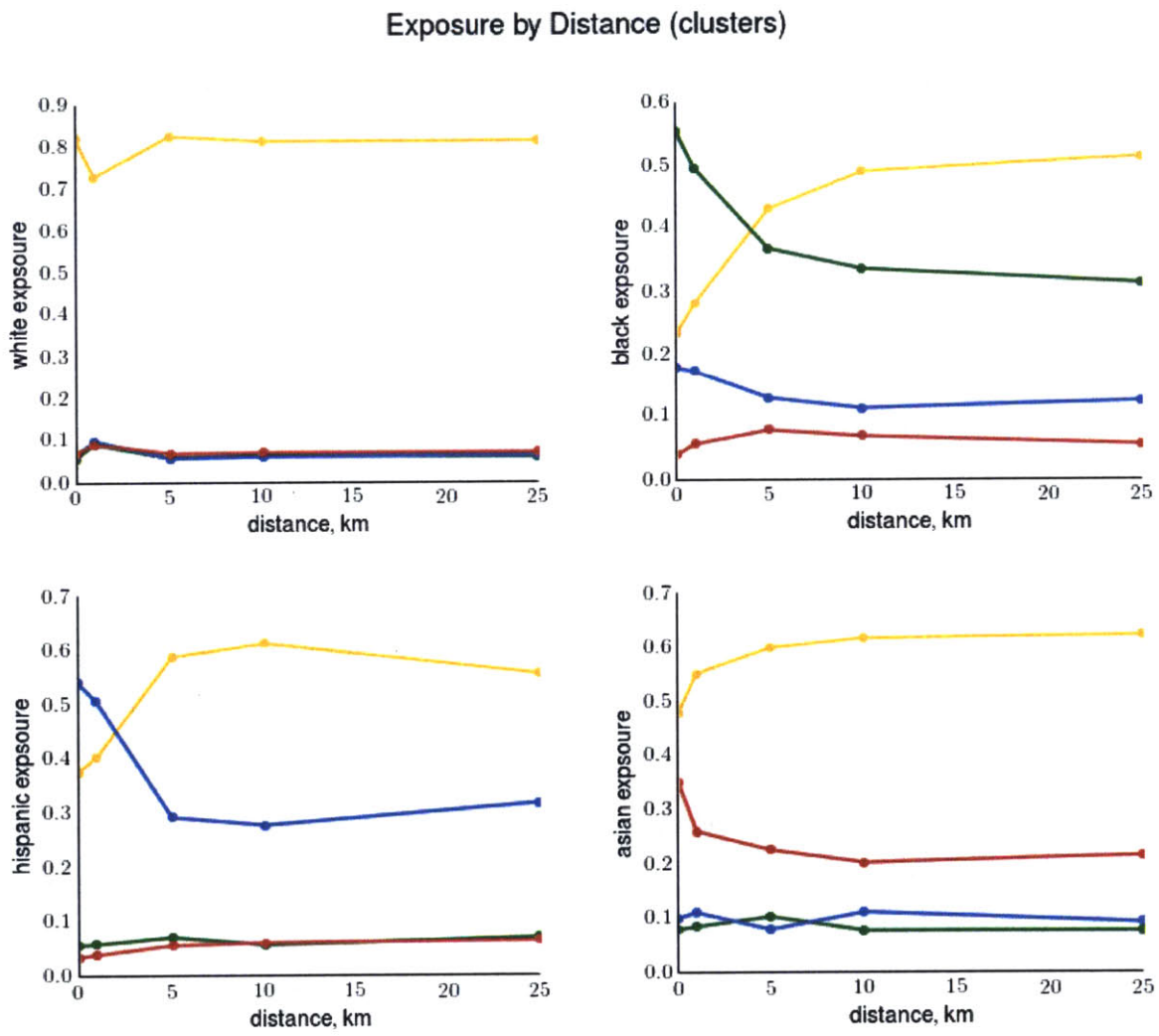


Figure 3-5: Each subplot represents the exposure for one race, in the top left we measure white exposure, the yellow line represents the exposure of white to white and the black line represents the exposure of white to black. etc. For all minorities as distance increases exposure to white increases and exposure to their own race decreases but remains higher than exposure to any other minority group

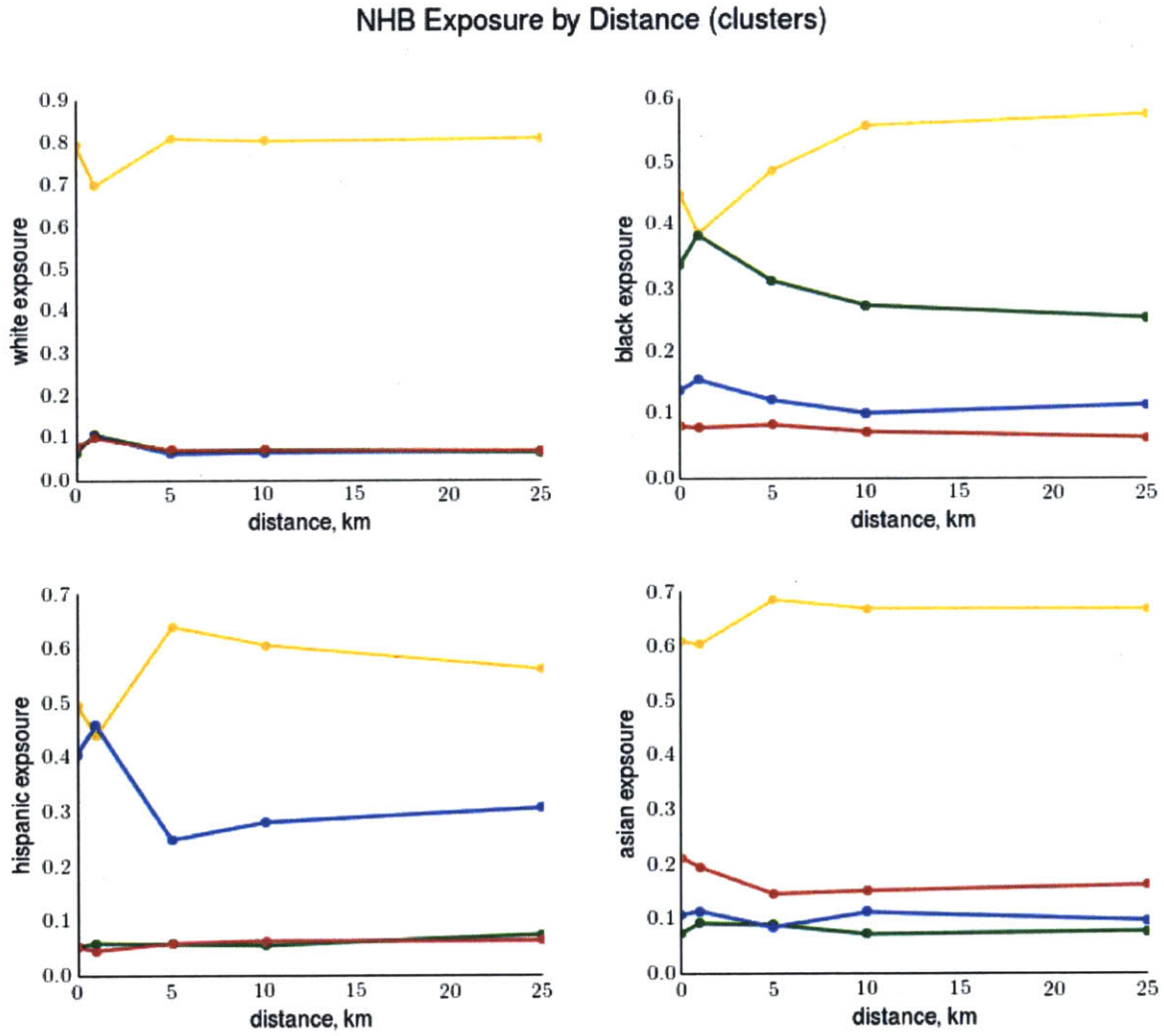


Figure 3-6: The same exposure metric as 3-5 calculated using the non home based OD network. Although the effects are slightly dampened when compared to 3-5 the same patterns are still apparent.

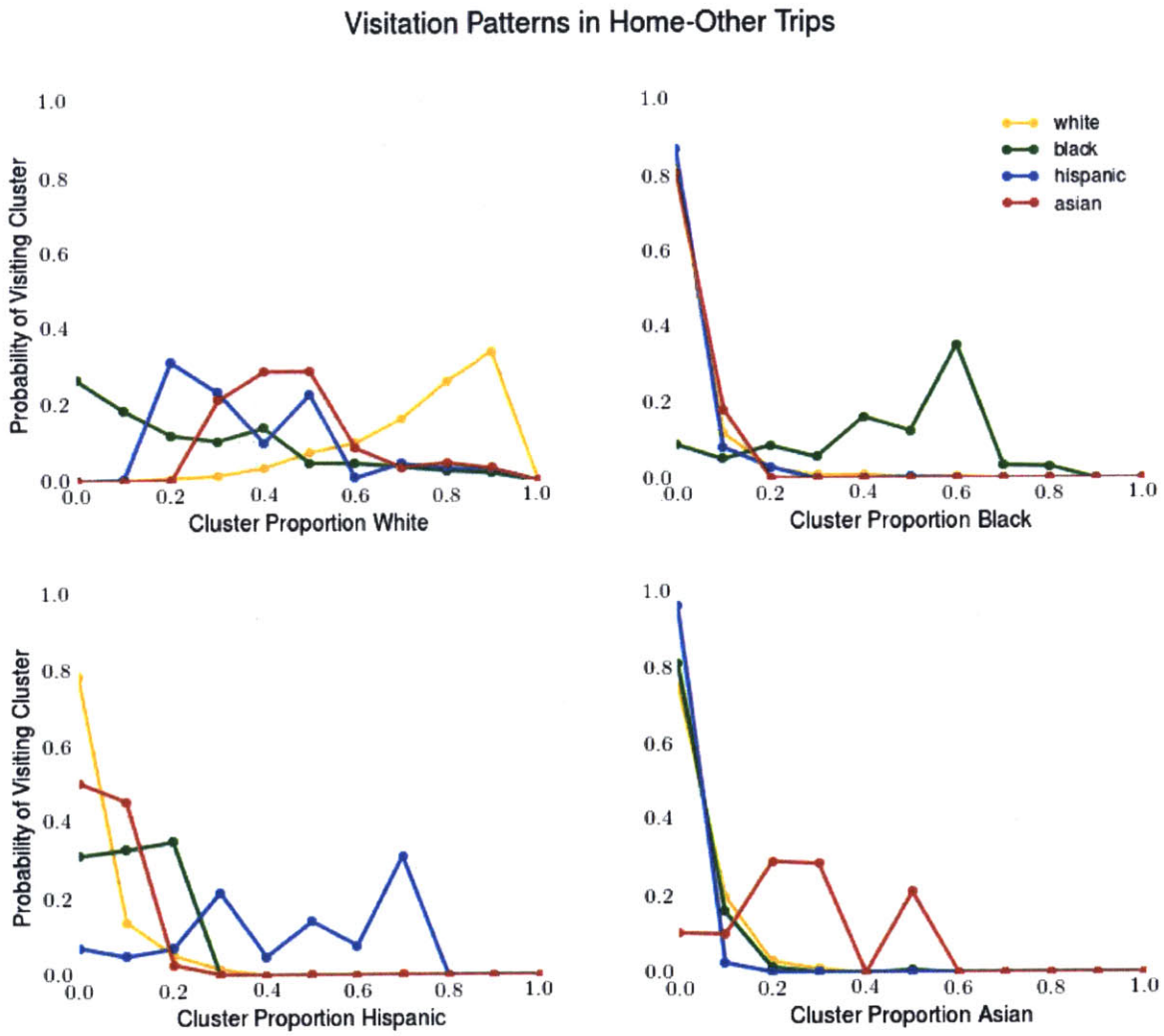


Figure 3-7: Visitation utility functions for each race

Chapter 4

Incorporating Race into a Mobility Model

4.1 Introduction

In describing collective movements the gravity model [3], radiation model [43] and variants of these aggregate models [2, 18] have been widely adopted to predict OD flows. The gravity model assumes that trips between origin and destination pairs decay with distance. The doubly constrained gravity model has been shown to preform well at many scales but grows in computational complexity as the number parameters increase whenever the number of zones in consideration grows. Furthermore, to properly to calibrate the parameters the model requires accurate input of the total trip production and the attraction volumes. The intervening opportunity model alternatively argues that trip volume decays with respect to the number of opportunities between an origin and destination. The radiation model, inspired by the intervening opportunity model, uses population density as an approximation of opportunities. It offers remarkable flexibility and simplicity due to its closed analytical form of the t_{ij} distribution written below:

$$t_{ij} = t_i \frac{P_i P_j}{(P_i + P_{ij})(P_i + P_{ij} + P_j)} \tag{4.1}$$

Where t_{ij} is the flow from location i to j and P_i is the population at i and P_{ij} is the population in all locations closer to i than j, excluding both i and j. Thus, the radiation model implicitly incorporates the heterogeneity in population distributions when predicting flows. Nevertheless the performance of both models have been shown to vary in different scenarios [31]. Here we use a parameter free rank based model presented in [27] that has been validated at the intra-urban and inter-urban scale. This model uses spatial population distributions, similar to the radiation model, to compute ranks for locations and predicts trips inversely proportional to the rank of a location pair. The radiation model uses the same data but has been shown to lose accuracy at the intra-urban scale [51]. For our purposes this represents the perfect model to test whether the visitation patterns 3-7 and exposure stability of minorities to their own race regardless of distance 3-5 is simply due to the spatial patterning of the minority populations. In order words we want to test if these results are simply the effects of residential segregation, or whether they stem from a minority preference to visit locations that are similar to their home locations with respect to racial composition.

4.2 Model description

We utilize the rank model presented in [27] which provides a simple and stable formulation to predict mobility fluxes at several scales and implicitly geocodes geographic population density information, providing a powerful framework to test different variations of population. the implicit geography considerations of the model also form to make our analysis inherently spatial and not subject to the problem of scale. Thus we are afforded a null model. If total population is able to match the mobility patterns arising from all different segregated areas, that shows that the exposure and visitation patterns are simply due to the geographic nature of the residential segregation in the

area, whereas if using only a minority's population to predict their own trips we could assume individuals are moving with some preference. The latter also encodes spatial information on the distribution of minorities.

The model is formulated from the perspective of locations, where the likelihood of a trip from origin o to destination l is a rank function depend by the population that lies within a circle centered on o with radius of the distance between o and l denoted $h(o, l)$ the haversine distance function.

$$rank_o(l) = \sum_{i \in S} P_i \tag{4.2}$$

Where S is the set of all locations closer to o than l that is neither o nor l. So the rank is simply the P_{ol} from the radiation model. Letting $t(q)$ be the total number of trips that occur with the same rank q, the ratio $P(q)$ can be defined as:

$$P(q) = \frac{t(q)}{\sum\limits_{rank_i(j)=q} P_i P_j} \tag{4.3}$$

The denominator is the total number of pairs of people with location rank q, or all the people that could possibly make a trip with rank q, they find the probability of a trip with rank $P(q) \propto \frac{1}{q}$. Since the the probability of a trip with rank q is inversely proportional to the rank itself, rank might be more suitable than geographical distance to characterize human movement. In terms of the above equation the rank based mobility model is formulated as such:

$$\frac{t_{ij}}{P_i P j} = \frac{t_i P(j|i)}{P_i P j} \propto \frac{1}{rank_i(j)} \tag{4.4}$$

Then $P(j|i)$ is normalized to:

$$P(j|i) = \frac{\frac{P_j}{rank_i(j)}}{\sum_{k \neq i} \frac{P_k}{rank_i(k)}} \tag{4.5}$$

When the locations are fine grained and there are a relatively high number of them the denominator can be approximated by:

$$\sum_{k \neq i} \frac{P_k}{rank_i(k)} \approx \sum_{n=1}^{M-1} \frac{1}{n} \approx \ln M \tag{4.6}$$

Where M is the total population. This allows us to model the flow t_{ij} as:

$$t_{ij} = t_i \frac{\frac{P_j}{rank_i(j)}}{\sum_{k \neq i} \frac{P_k}{rank_i(k)}} \approx t_i \frac{P_j}{rank_i(j) \ln M} \tag{4.7}$$

In relation to the radiation model:

$$t_{ij} \propto \frac{P_j}{rank_i(j)^2} \tag{4.8}$$

and in the case of this rank model:

$$t_{ij} \propto \frac{P_j}{rank_i(j)} \tag{4.9}$$

Since the radiation model prefers trips with smaller ranks, meaning shorter distances, this could explain why the radiation model is not suited for intra-urban trip prediction but rather longer range commuting fluxes.

We apply the rank model in two cases, one where we use the total population P_i of each location i to calculate ranks and predict t_{ij}^r for each race r. In the second

version we use only the population of race r, P_{ir} of each location i to calculate ranks and predict t^r_{ij} for each race.

4.3 Results

To evaluate the performance of both models we use a metric based on the Sorensen index which is used to quantify what extent the model predictions can reproduce the empirical t_{ij} distribution. The measure is defined as:

$$SSI = \frac{2\sum_i \sum_j \min(t_{ij}, t'_{ij})}{\sum_i \sum_j t_{ij} + \sum_i \sum_j t'_{ij}} \tag{4.10}$$

where t_{ij} represents the real OD pair and t'_{ij} represents the models prediction. The measure of 1 means complete equality and 0 means total disagreement. In 4.1 we evaluate the performance of both models against the empirical race networks when predicting trips within segregated areas. We see that the race aware model does significantly better than the race blind model in all minority cases.

	Race Aware Model (SSI)	Race Blind Model (SSI)
white	.68	.66
black	.63	.47
hispanic	.58	.47
asian	.36	.08

Table 4.1: race aware and race blind model performance. The race aware model does much better in predicting minority fluxes

4.4 Discussion

Prediction based on the rank based models only depends on the population of locations and their relative rank, indicating that collective human mobility is largely driven by geographical distribution of population. Nevertheless, the larger increase in perfor-

mance of the race aware model would indicate that it is not the simply total population which matters for prediction the mobility fluxes of minorities. By incorporating racial demographic information in mobility prediction we show a significant improvement in model performance, meaning not only do we organize residentially according to racial preferences, we move according to these as well.

Chapter 5

Conclusions

Racial segregation in residential areas is an enduring social phenomenon that has been extensively studied with respect to both its causes and effects. Since then, sociologists have characterized the dramatic impact of segregation in neighborhoods on the economic outcomes of residents: specifically, it can lead to cultural isolation, concentrated disadvantage, heightened exposure to criminality, and reduced access to resources and opportunities More recently, economists have marshaled evidence to demonstrate how neighborhoods, their economic opportunities and social structure are absolutely critical in the economic success of children who grow up in them.

These studies clearly demonstrate the dramatic influence of neighborhoods, yet contemporary characterizations of neighborhoods rely on static descriptions of populations constructed from survey and census data. Although census figures can show remarkable changes in the racial compositions over long periods of time, they provide only a glimpse of neighborhood exposure since they fail to capture any dynamic aspects of an individual's experience. Essentially, census data characterizes only where people sleep, not where they work or visit. As access to and efficiency of transportation in urban areas increases it has allowed for individuals move freely and municipal borders alone have become less meaningful indicators of where people spend their time. So incorporating mobility choices into metrics assessing exposure and neighborhood effects is increasingly important. Here we leverage daily trajectories of Boston residents to better understand the environments where individuals spend their time.

In this work we develop a unique platform to asses the socio demographic characteristics of the locations minorities visit. We do so by firstly cleaning, treating, and transforming noisy CDR data into meaningful and validated OD trips by purpose. The assignment of purpose allows to consider different trip types when characterizing individual's destination choices. We choose to study home based other trips because we believe they are mostly likely to be motivated by individual choice, while also retaining individual's home locations where people spend most of their time. Nevertheless, the framework we develop can be applied to any trip type, and in future studies we should consider all trips to get an accurate portrait of users' complete days.

Once we have created ODs that represent population scale travel patterns we create OD networks and use the topological property of node strength as input for reformulated metrics of entropy and exposure. By reformulate these measures to incorporate the racial compositions of the destinations users' visit, we are able to to better capture and characterize individual environments. The reformulated entropy metric reveals places of high mixing in the networks, or locations that are important in all race networks and the reformulated exposure metric shows that distance increases exposure to white but exposure to one's own race remains relatively constant after 5km.

Finally, we utilize two rank based models that implicitly incorporate geography and population distributions to explicitly see the effects of residential segregation on mobility patterns. We find the race blind model, where the total population is used to calculate the ranks and predict flows, does not accurately reproduce minority trips within segregated areas. We demonstrate how incorporating racial preferences in mobility choices can recover empirical mobility patterns, implying that many home other trips are induced by racial preference.

These simple models provide a very powerful platform to demonstrate that geographic patterning of residential segregation has a stark effect on what minorities are exposed to during the day but also reveals that geography is not the only factor at play. The significant increase in prediction accuracy of our race aware model demonstrates racial preference in mobility patterns. This work further add value to the mobility

modeling community, as we demonstrate a method to improve existing model and increase prediction accuracy. Taken together, this study demonstrates the potential and value of CDR data reaches far beyond simply predicting geographic movement, it can be used to consider human behavior from social and contextual perspective.

Bibliography

[1] Lauren Alexander, Shan Jiang, Mikel Murga, and Marta C. González. Validation of origin-destination trips by purpose and time of day inferred from mobile phone data. *submitted to Transportation Research Part C*, 2014.

[2] Duygu Balcan, Vittoria Colizza, Bruno Gonçalves, Hao Hu, José J Ramasco, and Alessandro Vespignani. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences*, 106(51):21484–21489, 2009.

[3] Marc Barthélemy. Spatial networks. *Physics Reports*, 499(1):1–101, 2011.

[4] Rahul C Basole. The value and impact of mobile information and communication technologies. In *Proceedings of the IFAC Symposium on Analysis, Modeling & Evaluation of Human-Machine Systems*, pages 1–7, 2004.

[5] Richard A Becker, Ramón Cáceres, Karrie Hanson, Ji Meng Loh, Simon Urbanek, Alexander Varshavsky, and Chris Volinsky. Clustering anonymized mobile call detail records to find usage groups. In *Ist Workshop on Pervasive Urban Applications*, 2011.

[6] Christopher Berry. Land use regulation and residential segregation: does zoning matter? *American Law and Economics Review*, 3(2):251–274, 2001.

[7] Dirk Brockmann, Lars Hufnagel, and Theo Geisel. The scaling laws of human travel. *Nature*, 439(7075):462–465, 2006.

[8] Jan K Brueckner, Jacques-Francois Thisse, and Yves Zenou. Why is central paris rich and downtown detroit poor?: An amenity-based theory. *European Economic Review*, 43(1):91–107, 1999.

[9] U.S. Census Bureau. *American Community Survey Design and Methodology*. U.S. Census Bureau, Washington, DC, 2009.

[10] David J Crandall, Lars Backstrom, Dan Cosley, Siddharth Suri, Daniel Huttenlocher, and Jon Kleinberg. Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences*, 107(52):22436–22441, 2010.

[11] David M Cutler, Edward L Glaeser, and Jacob L Vigdor. When are ghettos bad? lessons from immigrant segregation in the united states. *Journal of Urban Economics*, 63(3):759–774, 2008.

[12] Manlio De Domenico, Antonio Lima, and Mirco Musolesi. Interdependence and predictability of human mobility and social interactions. *Pervasive and Mobile Computing*, 9(6):798–807, 2013.

[13] Nancy A Denton and Douglas Massey. American apartheid: Segregation and the making of the underclass, 1993.

[14] Pierre Deville, Catherine Linard, Samuel Martin, Marius Gilbert, Forrest R Stevens, Andrea E Gaughan, Vincent D Blondel, and Andrew J Tatem. Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences*, 111(45):15888–15893, 2014.

[15] Robert D Dietz. The estimation of neighborhood effects in the social sciences: An interdisciplinary approach. *Social Science Research*, 31(4):539–575, 2002.

[16] Nathan Eagle, Alex Sandy Pentland, and David Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, 106(36):15274–15278, 2009.

[17] Tini Garske, Hongjie Yu, Zhibin Peng, Min Ye, Hang Zhou, Xiaowen Cheng, Jiabing Wu, and Neil Ferguson. Travel patterns in china. *PloS one*, 6(2):e16364, 2011.

[18] Segun Goh, Keumsook Lee, Jong Soo Park, and MY Choi. Modification of the gravity model and application to the metropolitan seoul subway system. *Physical Review E*, 86(2):026102, 2012.

[19] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[20] Mark Granovetter. Threshold models of collective behavior. *American journal of sociology*, pages 1420–1443, 1978.

[21] Mark Granovetter and Roland Soong. Threshold models of diffusion and collective behavior. *Journal of Mathematical sociology*, 9(3):165–179, 1983.

[22] Avery M Guest, Charis E Kubrin, and Jane K Cover. Heterogeneity and harmony: Neighbouring relationships among whites in ethnically diverse neighbourhoods in seattle. *Urban Studies*, 45(3):501–526, 2008.

[23] Ramaswamy Hariharan and Kentaro Toyama. Project lachesis: parsing and modeling location histories. In *Geographic Information Science*, pages 106–124. Springer, 2004.

[24] Steven R Holloway, Richard Wright, and Mark Ellis. The racially fragmented city? neighborhood racial segregation and diversity jointly considered. *The Professional Geographer*, 64(1):63–82, 2012.

[25] Shan Jiang, Gaston A Fiore, Yingxiang Yang, Joseph Ferreira Jr, Emilio Frazzoli, and Marta C González. A review of urban computing for mobile phone traces: current methods, challenges and opportunities. In *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*, page 2. ACM, 2013.

[26] Barrett A Lee, Sean F Reardon, Glenn Firebaugh, Chad R Farrell, Stephen A Matthews, and David O'Sullivan. Beyond the census tract: Patterns and determinants of racial segregation at multiple geographic scales. *American Sociological Review*, 73(5):766–791, 2008.

[27] Xiao Liang, Jichang Zhao, and Ke Xu. A universal law in human mobility. *arXiv preprint arXiv:1401.3918*, 2014.

[28] Thomas Louail, Maxime Lenormand, Oliva García Cantú, Miguel Picornell, Ricardo Herranz, Enrique Frias-Martinez, José J Ramasco, and Marc Barthelemy. From mobile phone data to the spatial structure of cities. *arXiv preprint arXiv:1401.4540*, 2014.

[29] Douglas S Massey and Nancy A Denton. The dimensions of residential segregation. *Social forces*, 67(2):281–315, 1988.

[30] Douglas S Massey and Jonathan Rothwell. The effect of density zoning on racial segregation in us urban areas. *Urban Affairs Review*, 2009.

[31] A Paolo Masucci, Joan Serras, Anders Johansson, and Michael Batty. Gravity versus radiation models: On the importance of scale and heterogeneity in commuting flows. *Physical Review E*, 88(2):022812, 2013.

[32] Philip McCann. *Modern urban and regional economics*. Oxford University Press, 2013.

[33] Massachusetts Department of Transportation. Massachusetts travel survey, 2012.

[34] Robert E Park. The urban community as a spatial pattern and a moral order. *The urban community*, pages 3–18, 1926.

[35] Carlo Ratti, Stanislav Sobolevsky, Francesco Calabrese, Clio Andris, Jonathan Reades, Mauro Martino, Rob Claxton, and Steven H Strogatz. Redrawing the map of great britain from a network of human interactions. *PloS one*, 5(12):e14248, 2010.

[36] Sean F Reardon, Stephen A Matthews, David OâĂŹSullivan, Barrett A Lee, Glenn Firebaugh, Chad R Farrell, and Kendra Bischoff. The geographic scale of metropolitan racial segregation. *Demography*, 45(3):489–514, 2008.

[37] Sean F Reardon and David OSullivan. Measures of spatial segregation. *Sociological methodology*, 34(1):121–162, 2004.

[38] Camille Roth, Soong Moon Kang, Michael Batty, and Marc Barthélemy. Structure of urban movements: polycentric activity and entangled hierarchical flows. *PloS one*, 6(1):e15923, 2011.

[39] Robert J Sampson and W Byron Groves. Community structure and crime: Testing social-disorganization theory. *American journal of sociology*, pages 774–802, 1989.

[40] Robert J Sampson, Jeffrey D Morenoff, and Thomas Gannon-Rowley. Assessing" neighborhood effects": Social processes and new directions in research. *Annual review of sociology*, pages 443–478, 2002.

[41] Thomas C Schelling. Dynamic models of segregation. *Journal of mathematical sociology*, 1(2):143–186, 1971.

[42] Christian M Schneider, Vitaly Belik, Thomas Couronné, Zbigniew Smoreda, and Marta C González. Unravelling daily human mobility motifs. *Journal of The Royal Society Interface*, 10(84):20130246, 2013.

[43] Filippo Simini, Marta C González, Amos Maritan, and Albert-László Barabási. A universal model for mobility and migration patterns. *Nature*, 484(7392):96–100, 2012.

[44] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[45] Lijun Sun, Kay W Axhausen, Der-Horng Lee, and Xianfeng Huang. Understanding metropolitan patterns of daily encounters. *Proceedings of the National Academy of Sciences*, 110(34):13774–13779, 2013.

[46] Christian Thiemann, Fabian Theis, Daniel Grady, Rafael Brune, and Dirk Brockmann. The structure of borders in a small world. *PloS one*, 5(11):e15422, 2010.

[47] Jameson L Toole, Serdar Colak, Fahad Alhasoun, Alexandre Evsukoff, and Marta C Gonzalez. The path most travelled: Mining road usage patterns from massive call data. *arXiv preprint arXiv:1403.0636*, 2014.

[48] Ott Toomet, Siiri Silm, Erki Saluveer, Tiit Tammaru, and Rein Ahas. Ethnic segregation in residence, work, and free-time: Evidence from mobile communication. *University of Tartu*, 2012.

[49] Duncan J Watts, Roby Muhamad, Daniel C Medina, and Peter S Dodds. Multiscale, resurgent epidemics in a hierarchical metapopulation model. *Proceedings of the National Academy of Sciences of the United States of America*, 102(32):11157–11162, 2005.

[50] Gregory R Weiher. Public policy and patterns of residential segregation. *The Western Political Quarterly*, pages 651–677, 1989.

[51] Yingxiang Yang, Carlos Herrera, Nathan Eagle, and Marta C González. Limits of predictability in commuting flows in the absence of data for calibration. *Scientific reports*, 4, 2014.