Human Movement Patterns of Different Racial-Ethnic and Economic Groups In U.S. Top 50 Cities: What Can Social Media Tell Us About Segregation?

Meiliu Wu

Advisor: Qunying Huang

A thesis submitted in partial fulfillment of

the requirements for the degree of

Master of Science

Cartography and Geographic Information Systems

at the

University of Wisconsin-Madison

2019

Acknowledgements

The completion of this Master thesis would have been impossible without the tremendous help and guidance from a group of people, to whom I am sincerely grateful. First of all, I would like to express my deepest appreciation to my advisor/role model Dr. Qunying Huang, who always gives me dedicated advice and support in both my life and work, especially at my hardest times. Her brilliant inspiration, strong enthusiasm, and heart-warming encouragement keenly touched me and led me through this tough project and beyond.

I am also very thankful to my committee members Dr. David Wong, Dr. Robert Roth, and Dr. Song Gao, who provided me with crucial advice and feedback for revision, and to Dr. Lisa Naughton who earnestly advised me on drafting the proposal.

I would like to also thank my supportive, lovely, and kind colleagues. Thank you to Xinyi Liu for offering me her computing facilities and helping me solve whatever programming bugs I struggled with, to Bo Peng for discussing with me all kinds of tricky problems and abstruse theories/algorithms in machine learning, and to Chenxiao (Atlas) Guo for sharing the office with me and working together in our final project as well as TAing the same class. Additionally, thanks to Clara Risk for proofreading my proposal, to Yuhao Kang for providing me with many stimulating ideas in research, to Yunlei Liang for being a caring and cheerful classmate/neighbor, and to all other geograds who provided a pleasant yet invigorating atmosphere for work and fun.

Finally, I would like to thank my dearest families and friends who keep encouraging me whenever I feel weak, supporting me with whatever I need, and loving me deeply for who I am. I am so lucky to have you all.

Table of Contents

Abstract

Human movement patterns, one of the important fields in human mobility study, is significant for various practical applications, and many studies have proven that it is strongly impacted by individual socioeconomic and demographic background. On the other hand, social media has become more and more popular in studying human movement patterns because of its exclusive advantages compared with traditional data source (e.g., travel diary and American Community Survey from U.S. Census). While many efforts have been made on exploring the influences of age and gender on movement patterns using social media, this study aims to analyze and compare the movement patterns among different racial-ethnic and economic groups using social media (i.e., geotagged tweets) from the U.S. top 50 populated cities. Since individual racial-ethnic information is usually not revealed by the social media users, we adopt different name-ethnicity prediction models to infer the race-ethnicity for the users. As for the individual economic status inference, the median house value from U.S. Census is utilized as a reliable reference.

Results show that the average tweeting density (i.e., tweeting frequency divided by the population of the group) of rich groups is 17% lower than the one from poor groups, but for different racial-ethnic groups with the same tweeting density, our results reveal that Non-Hispanic Black or African American have 5% more activity zones than Non-Hispanic Two or More Races and Hispanic or Latino origin. As for median travel distance, poor groups travel 42% shorter than rich groups. On the other hand, the median travel distance of Non-Hispanic White is 23% longer than the one of Hispanic or Latino origin, 18% longer than the one of Non-Hispanic Two or More Races, 10% longer than the one of Non-Hispanic Asian and Native Hawaiian and Other Pacific Islander, 9% longer than the one of Non-Hispanic American Indian and Alaska Native. To explore further on outbound-city travels (i.e., the travels with origins inside the boundary of the Urban

Areas defined by the U.S. Census but with destinations outside of that boundary), poor groups contribute 10% less outbound-city travels than rich groups. Particularly, the poor groups from the racial-ethnic minorities such as Non-Hispanic American Indian and Alaska Native (18%), Non-Hispanic Asian and Native Hawaiian and Other Pacific Islander (16%), and Hispanic or Latino origin (14%) have much lower percentages, while the poor groups from Non-Hispanic White and Non-Hispanic Black or African American reach the overall mean percentage (25%). This finding strongly proves that people who are economically disadvantaged and racial-ethnic minorities are more restricted in long distance travels, which indicates their spatial mobility is more limited into the local scale.

Another important finding is the economically-segregated movement pattern in the national scale – rich neighborhoods are mostly visited by the rich, while poor neighborhoods are mainly accessed by the poor, but some race-ethnicity groups can diversify this segregated pattern in the local scale, such as the Non-Hispanic Asian and Native Hawaiian and Other Pacific Islander + Poor group in New York having a much higher percentage (38%) of traveling to rich community than the national average level of p-to-r travels (12%).

Lastly, spatial variability of travel distances is also revealed. Although there is a uniform pattern of travel distance distributions among the U.S. top 50 populated cities, which are fitting a decreasing curve with long-tail, yet the median travel distance for the top 6 cities are significantly different (e.g., New York City 6245 m; Los Angeles 7362 m; Chicago 6807 m; Houston 9729 m; Philadelphia 7233 m; and Phoenix 8827 m). On the other hand, for the percentage of outbound-city travels of the U.S. top 6 cities, it shows that New York (27.2%) and Houston (27.5%) have more outbound-city travels, which could indicate their lightly stronger interaction power with other cities, while Los Angeles (22.8%) and Chicago (21.7%) have less outbound-city travels.

Chapter 1. Introduction

1.1. Background

Human mobility study is an important research thrust in GIScience and significant for a broad range of applications, such as public health and disaster management (Song, Zhang et al. 2014). One of these research themes is to examine human movement patterns (Gonzalez, Hidalgo et al. 2008), which is boosted by several reasons. For example, the characteristics of human movement patterns strongly influence urban formation, evolvement, as well as future planning (Waddell 2002). Also, a better understanding of the movement patterns can assist in controlling the spreading of contagious diseases among certain groups of individuals (Longini, Nizam et al. 2005). Moreover, the collective movement analysis (i.e., movement analysis of one specific group of individuals) can not only provide insights in travel modeling (Leutzbach 1988), but also diagnose the traffic-related abnormal events in a much larger scale of area (e.g., city, nation-wide, and international level), which outperforms the computer-vision-based simulation that is utilized specifically for the crowd behavior analysis in a smaller, localized scale (Andrade, Blunsden et al. 2006, Mehran, Oyama et al. 2009, Peng, Jin et al. 2012).

Human movement patterns at different geographic scales are strongly related to the socioeconomic and demographic background (SDB) (Bagrow and Lin 2012, Huang and Wong 2016). Meanwhile, human mobility is also highly constrained by transport accessibility, commuting facilities, as well as urban spatial structure (Black and Conroy 1977). The reason is that different SDB groups implicate different adaptabilities to these constraints, resulting in diversified movement patterns. For example, Dong, Ben-Akiva et al. (2006) proposed an activity-based accessibility measure, which reflected that SDB (e.g., employment status, household income, gender, race-ethnicity, etc.) is significant in differing human accessibility, including the aspects of

the number of vehicles per household, transportation mode, house location preferences, etc. Also, Crane (2007) used panel data from the American Housing Survey to measure and explain commute trends for the entire United States from 1985 through 2005, and he found that men's and women's commuting distances converged only slowly and commuting times diverged. Additionally, Blumenberg and Shiki (2007) used data from the 2000 Public Use Microdata Sample (PUMS) of the U.S. Census to examine the commute mode choice of California's foreign-born population and, more specifically, the relationship between length of residency in the U.S. and transit usage rates, and their results revealed that recent immigrants - regardless of race-ethnicity - were significantly more likely to commute by transit than native-born adults; after the first five years in the U.S., assimilation to automobile use occurred across all immigrant groups; however, the rate of assimilation varied significantly by racial-ethnic group even controlling for income. Asian immigrants rapidly moved to automobile use while Hispanic immigrants remained more likely to use transit than native-born commuters even after 20 years in the U.S. The findings from this study suggest that racial-ethnic differences affect commute mode choice.

Traditionally, the data of research in human movement patterns based on different SDB groups are collected from surveys, travel diary, or carry-on GPS devices to represent general patterns of local or national population displacement (Marion and Horner 2007, Kwan 2008, Antipova, Wang et al. 2011, Chen, Shaw et al. 2011). Nowadays, social media has become a more popular data source to support the analysis of human mobility due to its exclusive advantages, such as its accessibility to the public, a large amounts of participants, high spatiotemporal resolutions, and global coverage (Hasan, Zhan et al. 2013, Huang and Wong 2016, Luo, Cao et al. 2016).

However, previous studies in human mobility using social media mostly focus on individual behavior analysis, e.g., detecting daily frequent trajectory (Simini, González et al. 2012).

Only a few of them reveal collective travel characteristics of different SDB groups (Candia, González et al. 2008). For example, Huang and Wong (2016) determined that the travel destinations for individual's international trips vary for different economic groups, including rich, middle and poor, which are categorized based on median household income and house, and they also indicated that racial-ethnic backgrounds of individuals could contribute to this diversity. Also, Longley, Adnan et al. (2015) revealed and compared the level of ethnic segregation in residence in London during the hours of the working week and at weekends using Twitter data. Specifically, they created ten ethnic classes as study groups, including White British, Indian, Black African, Chinese, etc., based on the 2011 U.K. Census ethnic group category. Recently, Wang, Phillips et al. (2018) found a highly consistent travel pattern (e.g., the average travel distance and the number of neighborhoods spread in the metropolitan region) across the neighborhoods with different race and income characteristics in the U.S. 50 largest cities using Twitter data.

1.2. Research Objective

Although race-ethnicity has been recognized as a relatively sensitive field for research, yet it is important to dig in deeper based on these findings, especially when few studies have explored the extent how social media data can help investigate human travel patterns and further explored the travel patterns of different racial-ethnic groups. In addition, previous work using social media data has not revealed the interaction between economic and racial-ethnic factors, and which factor can better explain the diverse patterns of different groups. To fill this research gap, this study aims **to reveal, analyze, and compare the movement patterns among different racial-ethnic and economic groups using social media data from the U.S. top 50 populated cities**.

1.3. Research Approaches

The main challenge of using social media for this research lies in the fact that SDB data for the users are not available. Therefore, this study will apply prediction models to draw inference on the SDB of users to overcome this limitation. The inference of the SDB information from the social media users includes the prediction of gender, age, race-ethnicity, and socioeconomic status (SES), etc. While much progress has been made to identify user's gender and age using first names and the language use based on the social media messages (Bagrow and Lin 2012), and SES can be inferred based on the median house value of the census tract where the twitter user's predicted home is located (Huang and Wong 2016), techniques to draw inference on race-ethnicity are limited (Duggan and Brenner 2013). Existing prediction models typically use surnames as the means to infer race-ethnicity (Lauderdale and Kestenbaum 2000, Tucker 2005, Ambekar, Ward et al. 2009), but prediction accuracy is not always robust (Treeratpituk and Giles 2012). This blind spot needs attention.

Accordingly, this study will first apply the surname-based probabilistic model (i.e., Census Model) to detect the race-ethnicity of social media users. Besides users' surnames, we will build the Latent Dirichlet Allocation (LDA) prediction model which integrates first names as well. Since the LDA algorithm is an unsupervised classification method, we will also construct the Supervised LDA (sLDA) model to see if the prediction accuracy will be improved. The sLDA will generate a race-ethnicity labeled individual collection by referencing the U.S. Census Bureau's *Frequently Occurring Surname* list in 2010, with each surname corresponding to its race-ethnicity probability distribution (Chang, Rosenn et al. 2010). These three prediction models will be validated and compared with ground truth, and the best model will be selected for subsequent race-ethnicity predication and movement analysis.

After predicting social media users' race-ethnicity, which belongs to one of the six classes according to the U.S. Census Bureau's *Frequently Occurring Surname* list in 2010 (Comenetz 2016): 1) Non-Hispanic White; 2) Non-Hispanic Black or African American; 3) Non-Hispanic Asian and Native Hawaiian and Other Pacific Islander; 4) Non-Hispanic American Indian and Alaska Native; 5) Non-Hispanic Two or More Races; 6) Hispanic or Latino origin, we will detect their activity zones with geo-tagged social media posts based on the spatial clustering method. Then, each activity zone's economic status, classified into three categories: rich, middle, or poor community, will be inferred by the median house value of the census tract that the zone is spatially-within. Next, the home location (i.e., one of the activity zones) of each user will be identified based on the movement patterns as well as geographic context (i.e., the land use types) among all activity zones. Finally, the economic status of the user will be determined based on the home activity zone.

With both race-ethnicity and economic status identified, each user can be grouped into one of the 18 racial-ethnic and SES groups (i.e., the cross-classification of 6 races-ethnicities and 3 economic statuses, such as Non-Hispanic White + Rich, Non-Hispanic White + Middle, and Non-Hispanic White + Poor), and all users from the same group will be aggregated to reveal the collective movement pattern of each group. Specifically, different statistics, including the average number of activity zones, the spatial variability and demographic differences of travel distances, as well as the urban mobility spread in different economic destinations, will be analyzed graphically and quantitatively to explore the discrepancies of the movement patterns among different groups.

1.4. Contributions

Five main contributions are achieved in this study: 1) the collective movement pattern of each racial-ethnic and SES group is examined in detail, which provides valuable insights of how race-ethnicity and economic status can contribute to the diversity of the collective movement patterns as well as how significant their influences can be; 2) the collective movement patterns among all racial-ethnic and SES groups are compared together, which in turn can reveal and analyze their differences in the number of activity zones, travel distances in both local and national scales, as well as the economically-segregated travel destinations; 3) the spatial variability of movement patterns among the U.S. top 50 populated cities is also displayed; 4) we synthesize a massive amount of social media data integrated with GIS data (e.g., the U.S. census tracts, land use/land cover information) to infer home locations and SES for the social media users in this study. The data processing techniques and comprehensive analysis framework shed light on how big social media data and GIS could jointly provide knowledge in human movement patterns; 5) our novel methodology for the analysis of collective movement patterns is applicable to social media data (e.g., geotagged tweets) and beyond, such as cell phone location data and travel diary with carry-on GPS devices. Therefore, the methodology workflow will be a significant prototype in future for exploring the collective movement patterns using any point-based data source.

Chapter 2. Literature Review

2.1. Individual Trajectory Mining Using Social Media

To identify the individual trajectory using social media data, existing studies start off detecting individual's activity zones, and location types (e.g., home, and work) that the individual frequently visits. In such studies, data are most often collected and analyzed at the daily level. For example, Huang, Cao et al. (2014) proposed an approach that relied on a spatial clustering method to determine major activity sites. Since these major sites are frequently visited, the spatiotemporal points recording the locations of tweets are likely clustered around these sites. Similarly, Huang and Wong (2016) used the density-based spatial clustering of applications with noise (DBSCAN) algorithm to create spatial clusters based on the geo-tags of tweets for each twitter user. In that case, a set of clusters were produced and became activity zones where the twitter user frequently visited and posted tweets. While some tweet locations were included to form the clusters, dispersed point locations associated with rare events or activities that deviated from the regular patterns were excluded. To explore individual trajectory from the online footprints, a representative point (e.g., the median centroid) for each activity zone will be identified and geo-located (Huang, Li et al. 2016). To infer the economic status of each activities zone, GIS data (e.g., median house values by U.S. census tracts) that is spatially related to the zone will be considered. As for predicting the home location for each twitter user, land use/land cover data indicating residential areas will be integrated in order to identify the residential areas.

After detecting the home location for each twitter user, all other activity zones can be delineated to represent movement patterns of the individual (Shen, Kwan et al. 2013). To further explore the activity space, Huang and Wong (2016) observed the activity zones (representative

locations), the distance between home and activity zones, and the standard deviational ellipse to summarize the spatial distribution in mobility analysis for each tweeter.

However, these studies primarily focus on examining movement behaviors and patterns at the individual level. Candia, González et al. (2008) described that many studies had unfolded the daily movement pattern based on the predicted daily trajectory, but few of them focused on the demographic group travel pattern. For example, Huang and Wong (2016) developed an approach to determine twitter users' home and work locations in order to examine the activity patterns of individuals. To infer the SES of individuals, they incorporated the median house value and household income from American Community Survey (ACS) data. Using Twitter data in Washington, DC, the activity patterns of twitter users in DC with different SESs were analyzed. Although they illustrated how individual travel pattern could be detected by identifying the spatiotemporal clustering zones in the travel trajectory, yet they mainly emphasized on exploring the collective movement patterns among different SES groups only. In another research about investigating and comparing the level of ethnic segregation in residence in London during the hours of the working week and at weekends using Twitter data, Longley, Adnan et al. (2015) created ten ethnic classes as study groups, including White British, Indian, Black African, Chinese, etc., based on the 2011 U.K. Census ethnic group category. They observed that activity patterns of those classified as White British, White Irish, and White Other ethnic groups were scattered throughout the whole city. However, for certain ethnic groups, different areas of high Tweeting activities could be identified. The most segregated group of Twitter users that their methods coded to ethnicity were the Bangladeshis, and their level of segregation relative to other groups intensified during weekdays and at weekends, when the contrast with Indian and Pakistani Twitter users was particularly apparent. It was also interesting that the Irish and White Other groups were

segregated in terms of residence, although this was much less apparent for both groups in terms of weekday locations, and this pattern was similar for the Chinese. Additionally, Wang, Phillips et al. (2018) found a highly consistent travel pattern (e.g., the average travel distance and the number of neighborhoods spread in the metropolitan region) across the neighborhoods with different race and income characteristics in the U.S. 50 largest cities using Twitter data. However, this finding contradicts with previous urban and transportation studies – the mobility of the economically disadvantaged population is more restricted (Murakami and Jennifer 1997, Giuliano 2005, Paulley, Balcombe et al. 2006). Therefore, more efforts should be made on revealing as well as verifying the collective movement patterns of different economic and racial-ethnic groups. To fill this gap, this paper will investigate the collective movement pattern discrepancies based on the groups with the combination of economic statuses and races-ethnicities.

2.2 Race-Ethnicity Inference

Race-ethnicity is an influential factor for contributing to the differences in travel characteristics. As an example, Järv, Müürisepp et al. (2015) created the activity spaces of individuals over different temporal frames using the Estonia cell phone data, and their research implied that the differences in the characteristics of activity spaces (sizes and numbers of locations) between the Estonia- and Russian-speaking populations might be associated with their work and residence locations as well as their races-ethnicities. For example, the Russian-speaking minority was found to visit 45% fewer districts than Estonians, and their activity locations were more spatially concentrated. The smaller extent of spatial mobility observed in the Russian-speaking minority may indicate lower social status in relation to their daily use of space, as well as more limited integration into society.

There is a rich history of authors using associations between surnames and races-ethnicities in curated sources such as census data to infer races-ethnicities, which is also known as the surname-based Census Model (Lauderdale and Kestenbaum 2000, Tucker 2005, Ambekar, Ward et al. 2009). However, surname analysis is not always accurate enough to identify specific racial-ethnic groups such as African Americans and Hispanics (Fiscella and Fremont 2006). Therefore, more advanced statistic methods have been also examined for the race-ethnicity inference, such as LDA model based on Gibbs sampling, which is a topic model based on the premise that each word in each document comes from a topic and the topic is selected from a per-document distribution over topics. However, traditional LDA is an unsupervised prediction model, and Ramage, Hall et al. (2009) shows that traditional LDA is outperformed by the sLDA model, which is a topic model that constrains LDA by defining a one-to-one correspondence between LDA's latent topics and user tags. Thus, the sLDA model solves the issue of LDA that topics are randomly assigned to each word for the initialization of Gibbs sampling. This allows sLDA to directly learn word-tag correspondences.

Later, Chang, Rosenn et al. (2010) proposed a lightly supervised LDA model using census surname data to predict the race-ethnicity of individuals based on their first names and last names. They demonstrated that their approach was able to predict the races-ethnicities of individuals better than the existing name-ethnicity prediction models. Predicting the race-ethnicity of each individual in a population also enabled them to understand the different demographic characteristics of each race-ethnicity. For example, they found that the statistics from the U.S. Census about different proportions of races-ethinicities would underestimate the one of Non-Hispanic Asians and Native Hawaiians and Other Pacific Islanders on Facebook, and overestimate the one of Non-Hispanic Black or African American users on Facebook. Also, they analyzed how social media usage,

language, political affiliation, gender, and geography depend on race-ethnicity. Similarly, Ambekar, Ward et al. (2009) built a classifier using hidden Markov models (HMMs) to classify first names and surnames into 13 cultural/ethnic groups with individual group accuracy comparable to earlier binary (e.g., Spanish/non-Spanish) classifiers. They applied the classifier to over 20 million names from a large-scale news corpus, identifying interesting temporal and spatial trends on the representation of particular cultural/ethnic groups.

Therefore, in order to gain the best performance of race-ethnicity inference for our social media users, this study will compare the surname-based Census Model, the LDA model, as well as the sLDA model. Our movement analysis will be based on the one that obtains the best race-ethnicity prediction results.

Chapter 3. Methodology

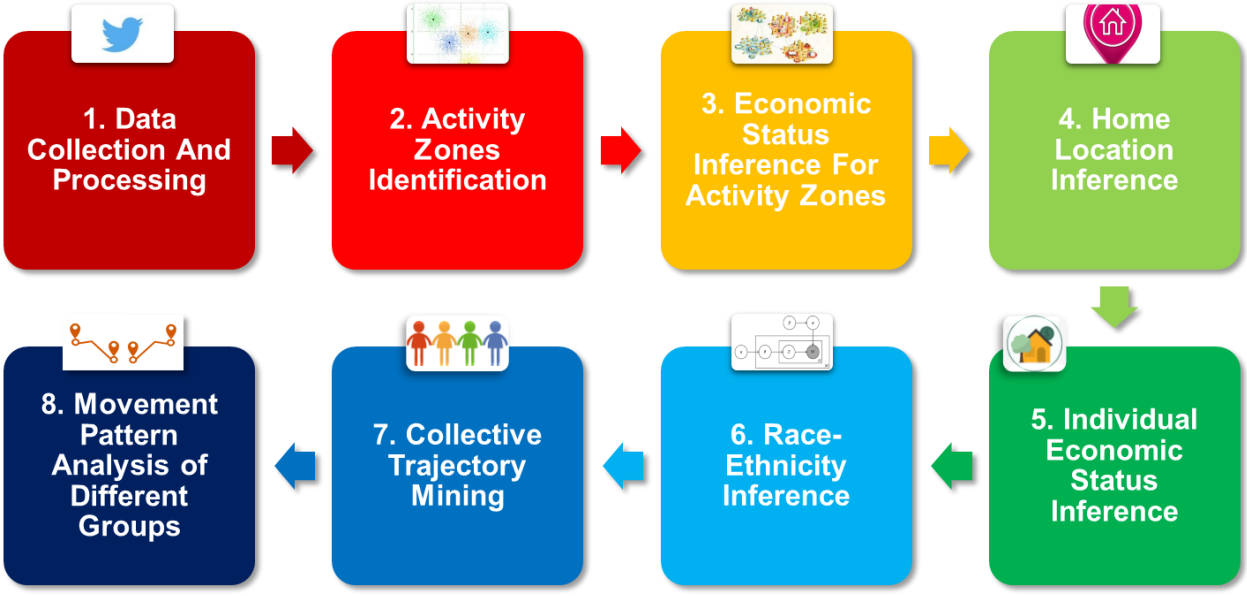


Figure 1. The workflow to collect, process and analyze Twitter data to explore the spatiotemporal characteristics of different economic and racial-ethnic groups

The workflow of this study is shown as Figure 1 above. This workflow consists of eight main procedures:

(1) Data Collection and Processing: The tweets are harvested globally via the Twitter's streaming application program interface (API). Then they are stored in PostgreSQL 10. Next, we process the data by filtering the tweets within the U.S. top 50 populated cities defined as Urban Areas by U.S. Census as well as filtering the twitter users with valid first names and last names for the subsequent race-ethnicity inference.

(2) Activity Zones Identification: Spatial clustering algorithm (i.e., DBSCAN) is applied to create spatial clusters for the geo-tagged tweet points, and those clusters are considered as the activity zones for each twitter user.

(3) Economic Status Inference for Activity Zones: The ACS data from U.S. Census is used as a reliable reference to infer the economic status of each activity zone. Specifically, we take account of the median house value by census tract and classify the census tracts into three categories of economic status: rich, middle, and poor.

(4) Home Location Inference: Each twitter user's home location is considered as the activity zone whose median centroid is located within residential areas as well as having the largest total number of daily inter-zone travels (i.e., the daily travels "into" and "out of" this activity zone) among all activity zones.

(5) Individual Economic Status Inference: The economic status of each twitter user is inferred based on the economic status of the predicted home location (see Section 3.3 and 3.4).

(6) Race-Ethnicity Inference: Three prediction models are applied to infer individual race-ethnicity using first names and last names, and the one with the best results will be utilized for the subsequent analysis of collective movement patterns.

(7) Collective Trajectory Mining: All twitter users with inferred race-ethnicity and economic status are aggregated into one of the 18 different groups (6 races-ethnicities x 3 economic statuses).

(8) Movement Pattern Analysis of Different Groups: For each group, we observe the average number of activity zones, the spatial variability and demographic differences of travel distances, as well as the urban mobility spread in different economic destinations.

3.1. Data Collection And Processing

In this research, we will use Twitter as the data source. Studies (Morstatter, Pfeffer et al. 2013) show that Twitter data are capable of providing snapshots of human daily trajectory at a macro scale (more than 500 million registered users publishing 400 million tweets per day in 2013), a coverage that is impossible for using the traditional survey approach. Therefore, using tweets as the social media data for human mobility research is tenable and justifiable.

The U.S. top 50 populated cities were chosen as the study cases. Besides their large spatial coverage of the whole nation, they were selected due to three additional reasons. First, to highlight the significance of race-ethnicity on collective mobility patterns, the studied population should have sufficient variation in ethnicity. As population in racial-ethnic minority status is often under-represented, this population needs to be "oversampled". For example, Washington DC has 49% Non-Hispanic Black or African American, 43.6% Non-Hispanic White, 5.0% other (including Non-Hispanic American Indian and Alaska Native, Non-Hispanic Native Hawaiian and Other Pacific Islander), and 3.1% Non-Hispanic Asian; among these, there were 8.3% Hispanic or Latino origin and 1.6% Non-Hispanic Two or More Races according to the 2008 Census (estimates); Chicago also has diverse race-ethnicity composition based on the results of the 2010 Census, including 31.7% Non-Hispanic White, 32.9% Non-Hispanic Black or African American, 5.5% Non-Hispanic Asian and Native Hawaiian and Other Pacific Islander, 2.7% Non-Hispanic Two or More Races, 0.5% Non-Hispanic American Indian. Thus, these 50 cities should include enough ethnic-minority population. Second, most cities have relatively distinctive neighborhood characteristics. For example, in both Chicago and Washington DC, there is a northwest affluent quadrant and a southeast deprived quadrant. Since residential segregation has strong correlations with SES and race-ethnicity, such clearly fragmented urban landscape facilitates the analysis of

the roles of SES and race-ethnicity in affecting movement patterns. Third, most populated cities have a relatively larger population adopting ICT, including the use of Twitter. Thus, it is more likely to identify sufficient twitter users for our case study by using the U.S. top 50 populated cities. Only tweets with a geo-tag are used. Based on the predicted home locations inferred from the geo-tag tweets, we selected the twitter users who live in those 50 cities.

To collect the Twitter data, we first used the Twitter's streaming API to harvest geo-tagged tweets globally. Overall, more than 344 million geo-tagged tweets were collected for 18 months (from December 2013 to May 2015), among which over 110 million are located within the bounding box of the U.S. continent. Since we will infer race-ethnicity for each twitter user using first name and last name, we filtered the twitter users based on their user names and deleted the invalid ones that were identified as organizations (e.g., companies, advertisements) or provided insufficient information to be consider valid names (e.g., only one word in the user name field; only one letter in the first name or last name; containing non-English words). Finally, we identified ~1 million unique twitter users with valid first names and last names, who posted more than 37 million tweets within the U.S top 50 populated cities. These filtered tweets and twitter users will be used for the subsequent analysis.

3.2. Activity Zones Identification

To differentiate various movement patterns at an individual level, the locations that a twitter user visits equal to or more than 4 times can be considered as activity zones (Huang and Wong, 2016). To detect these locations, the geotagged tweets from each individual will be grouped as different spatial clusters (Huang, Cao et al. 2014). While some tweet points were included to form the

clusters, dispersed points associated with rare events or activities that deviated from the regular patterns were excluded.

As previous works (Huang and Wong 2016, Wang, Phillips et al. 2018), we will also apply DBSCAN algorithm to create the spatial clusters for each twitter user. DBSCAN needs two input parameters: 1) the upper bound distance (*eps*) for each point to be clustered into any existing group; 2) the minimum number of points (*minpts*)for a group to be considered as one valid cluster. In order to obtain a satisfying result of spatial clustering on our tweets, we referred to previous studies (Borah and Bhattacharyya 2004, Birant and Kut 2007, He, Tan et al. 2014) and conducted exploratory experiments using self-adjusted DBSCAN. The optimal result is obtained when the *eps* value is set as 50 meters, and the *minpts* as 4.

3.3. Economic Status Inference For Activity Zones

With spatial clusters (i.e., activity zones) detected from the geotagged tweets, the movement pattern can be explored at an individual level. Specifically, each activity zone will be represented by the median centroid of the zone, the point whose Euclidean distances to all other points in this zone sum up to be the minimum (Huang, Li et al. 2016).

The economic status of each activity zone is inferred by the economic status of the census tract that the median centroid of this zone is spatially-within. Specifically, the economic status includes poor, middle, and rich, which are classified corresponding to the median house values of all census tracts within a city – poor tracts are the ones with median house value less than the average median house value of the city minus one standard deviation, while rich tracts are the ones with median house value more than the average median house value of the city plus one standard

deviation, and middle tracts will be the ones in between (Huang and Wong 2016). Both census tracts and urban areas data were downloaded from ACS (5-year) 2010 - 2014.

For further exploration of the movement patterns among different racial-ethnic and economic groups, the number of tweets in each activity zone as well as the number of activity zones of each individual will also be recorded.

3.4. Home Location Inference

After identifying the activity zones for each twitter user, we assume that an individual leaves and returns home more frequently than any other activity zones in a daily manner. Therefore, the home location of each twitter user will be considered as the activity zone whose median centroid is spatially within residential areas referring to the National Land Cover Database (NLCD) as well as having the largest total number of daily inter-zone travels (i.e., the daily travels “into” and “out of” this activity zone) among all activity zones (Huang and Wong 2016). Additionally, the cluster size (i.e., the number of tweets in an activity zone) will be utilized to break the tie.

To obtain the total number of daily in-out travels for each cluster, tweets are first sorted based on the local posted time for each day. Since each tweet has been labeled the cluster that it belongs to, each cluster will then be calculated the total in-times and out-times for each day. Finally, the total number of daily in-out travels for each cluster will be the sum of all days within the whole time span.

3.5. Individual Economic Status Inference

The economic status of each twitter user will be inferred based on the economic status of the predicted home location, which is determined with the methods described in Section 3.3 and 3.4.

3.6. Race-Ethnicity Inference

Three name-ethnicity prediction models will be applied to infer individual race-ethnicity information. First, many previous studies have associated surnames with races-ethnicities from curated sources such as census data to infer races-ethnicities (Lauderdale and Kestenbaum 2000, Tucker 2005, Ambekar, Ward et al. 2009). Therefore, this study will also integrate this surname-based Census Model as one empirical experiment.

Additionally, in order to add first name as another variable, we will utilize the general techniques in big data mining such as Latent Dirichlet Allocation (LDA) and Gibbs Sampling etc. (Porteous, Newman et al. 2008). Our LDA model, a unsupervised classification method of machine learning, is built based on the "LDA" R-package developed by Jonathan Chang in 2015, and it will be leveraged to calculate the race-ethnicity distribution for each twitter user using his/her first name and surname.

Specifically, LDA is a three-level hierarchical Bayesian model, in which each twitter user from the users' database is modeled as a finite mixture over an underlying set of races-ethnicities with corresponding probabilities. Each race-ethnicity is, in turn, modeled as an infinite mixture over an underlying set of words (i.e., first names and surnames) with corresponding probabilities as well. Therefore, in the context of name-ethnicity modeling, the race-ethnicity probabilities provide an explicit representation of a twitter user. The Efficient approximate inference techniques are presented based on variational methods and an Expectation–Maximization (EM) algorithm for empirical Bayes parameter estimation (Blei, Ng et al. 2003). Specifically, at the initialization stage of Gibbs Sampling, each word will be assigned randomly a race-ethnicity. Then, the model will calculate the number of each word existing in each race-ethnicity class (β_r^f or β_r^l) as well as the

number of each race-ethnicity class existing in each twitter user (θ_n^r). For each iteration, the current word (f_n or l_n) will be calculated a new set of probabilities for each race-ethnicity class based on all other words, which determines the new assignment of race-ethnicity to it. Then, the model will update β_r^f or β_r^l as well as θ_n^r for this twitter user n. Iterations of the algorithm will terminate until the race-ethnicity probability distribution of each twitter user n (i.e., θ_n^r) and the word probability distribution of each race-ethnicity class r (i.e., β_r^f and β_r^l) have reached convergence. The predicted results will include the probability distribution of each race-ethnicity class for all twitter users, as well as the probability distribution of words (i.e., names) for all race-ethnicity classes (Figure 2).

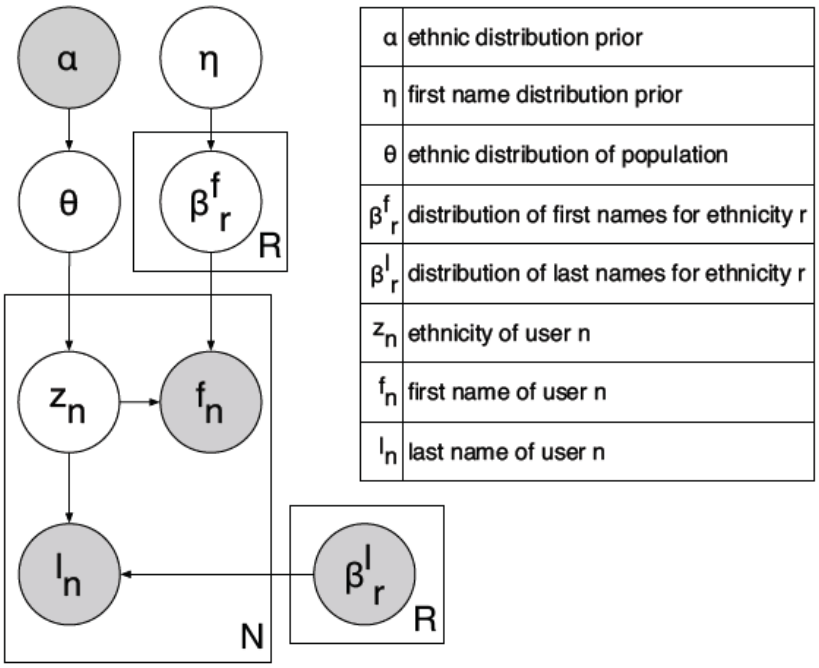


Figure 2. A graphical model representation of the model used to infer ethnicities. Shaded nodes are observed variables and unshaded nodes are unobserved. Plates indicate replication.

The study also involves sLDA model. The sLDA is a probabilistic model that describes a process for generating a labeled document collection. Similar to LDA, sLDA models each document as a mixture of underlying topics and generates each word from one topic. However, unlike LDA, sLDA incorporates supervision by simply constraining the topic model to use only those topics that correspond to a document's (observed) label set. Previous studies (Chang, Rosenn et al. 2010, Treeratpituk and Giles 2012) demonstrated that their sLDA model was able to predict the races-ethnicities of individuals using first names and last names, which provided a better prediction correctness better prediction correctness than the existing name-ethnicity prediction models.

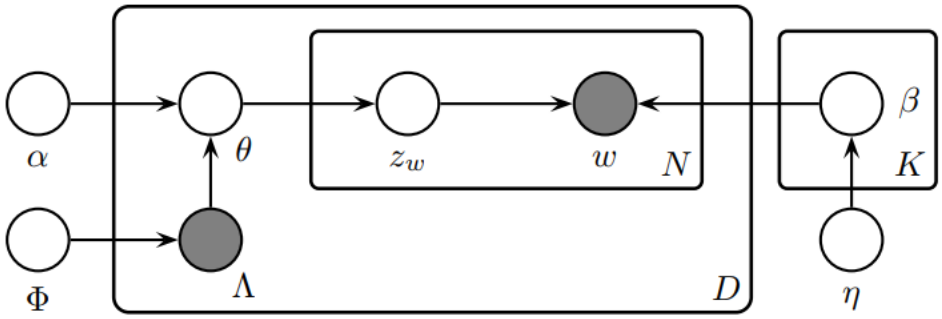


Figure 3. Graphical model of Supervised LDA. Unlike standard LDA, both the label set Λ as well as the topic prior α influence the topic mixture θ.

Thus, this study will compare the results from the surname-based Census Model, traditional LDA, as well as the sLDA. Result validation utilizes the ground truth dataset from 297 twitter users, which were obtained by manually identifying the apparent ethnicity based on the profile

photo of each twitter user. For the subsequent analysis, the twitter users who are successfully inferred race-ethnicity as well as economic status will be the final targeted population.

3.7. Collective Trajectory Mining

Individual trajectories will be aggregated based on the predicted group, i.e., one of the cross-classification of race-ethnicity and economic status groups. Specifically, this study includes six race-ethnicity classes, i.e., 1) Non-Hispanic White; 2) Non-Hispanic Black or African American; 3) Non-Hispanic Asian and Native Hawaiian and Other Pacific Islander; 4) Non-Hispanic American Indian and Alaska Native; 5) Non-Hispanic Two or More Races; 6) Hispanic or Latino origin, by referring to the classification from the U.S. Census Bureau's *Frequently Occurring Surname* list in 2010 (Comenetz 2016). The economic status includes poor, middle, and rich (Huang and Wong, 2016), which corresponds to the economic status of the twitter users. With the cross-classification of economic status and race-ethnicity, we will examine 18 (6 races-ethnicities x 3 economic statuses) different groups in total for this study.

3.8. Movement Pattern Analysis Of Different Groups

After detecting the home location for each twitter user, all other activity zones can be delineated to represent movement patterns of the individual (Shen, Kwan et al. 2013). We can use ArcGIS Pro as the visualization tools to show the activity zones.

One of the significances of the research is to detect the differences of movement patterns among different racial-ethnic and economic groups. Many studies have unfolded the daily movement pattern based on the predicted daily trajectory, but few of them realize the importance of the collective travel pattern (Candia, González et al. 2008). In this paper, we will fill this gap

by investigating the exploration of movement pattern discrepancies based on economic status and races-ethnicities.

As mentioned in Section 3.6, we will examine 18 different groups in total for this study. First, at the exploratory stage, the destination distributions for each group can be mapped out vividly by aggregating twitter users into groups. For a more statistical exploration and comparison of the movement patterns among different racial-ethnic and economic groups, we will observe the average number of activity zones, the spatial variability and demographic differences of travel distances, as well as the urban mobility spread in different economic destinations:

- **Average number of activity zones for different groups:** After processing the tweets, the number of activity zones of each twitter user has been calculated. In order to compare the activity diversities among different racial-ethnic and economic groups, the average number of activity zones for each group can be considered to represent the variation of activity diversities among different groups.
- **Spatial variability and demographic differences of travel distances:** Travel distances in different areas can reveal spatial variability. In this study, each one of the U.S. top 50 populated cities will be examined the travel distance distribution (i.e., proportion) based on the residents' median travel distances from home to all other activity zones. Similarly, the travel distances associated with a certain group can indicate the geographical extent of the group's activity space in general, which can be indicated by the travel distance distribution for each group. Specifically, we will further explore the spatial variability and demographic differences in longer travel distances as well. For example, the proportion of outbound-city travels out of all travels for each city and each group can be calculated and graphically visualized.

- **Urban mobility spread in different economic destinations:** For the inner-city travels (i.e., the travels within the boundaries of the Urban Areas defined by the U.S. Census), we will first visualize the economically-segregated movement pattern in a national level using the U.S. top 50 populated cities. Then we will examine in depth for each group in both New York and Los Angeles, since these two largest cities can provide more data for the race-ethnicity minorities. Specifically, to reveal the urban mobility spread in different economic destinations, we will calculate the proportion of each group that has access to poor, middle, or rich community respectively.

Chapter 4. Results and Analysis

4.1. Ethnicity Prediction Results and Validation

Table 1. Prediction results of the Supervise LDA model

Supervised LDA	Non-Hispanic White	Non-Hispanic Black or African American / Hispanic or Latino origin	Non-Hispanic Asian and Native Hawaiian and Other Pacific Islander / Non-Hispanic American Indian and Alaska Native	Non-Hispanic Two or More Races
Precision	0.754	0.883	0.845	N/A
Recall	0.636	0.704	0.781	N/A
F1 Score	0.690	0.784	0.812	N/A

Table 2. Prediction results of the surname-based Census Model

Census Model	Non-Hispanic White	Non-Hispanic Black or African American	Hispanic or Latino origin	Non-Hispanic Asian and Native Hawaiian and Other Pacific Islander	Non-Hispanic American Indian and Alaska Native	Non-Hispanic Two or More Races
Precision	0.924	0.894	0.960	0.974	0.983	N/A
Recall	0.677	0.881	0.810	0.947	1.000	N/A
F1 Score	0.781	0.887	0.879	0.961	0.991	N/A

In order to select the best model for individual ethnicity prediction, we constructed three models and compared their prediction results, including LDA, sLDA, and the Census Model. The validation utilizes the ground truth dataset from 297 twitter users, which were obtained by manually identifying the apparent ethnicity based on the profile photo of each twitter user, and specifically, we randomly collected 99 Non-Hispanic Whites, 67 Non-Hispanic Black or African Americans, 57 Non-Hispanic Asians and Native Hawaiians and Other Pacific Islanders, 16 Non-Hispanic American Indian and Alaska Natives, and 58 Hispanics or Latinos origin from the users' database. The metrics for validation include precision (i.e., the ratio of correctly predicted positive observations to the total predicted positive observations, which equals to TP/(TP+FP)), recall (i.e., the ratio of correctly predicted positive observations to the all observations in actual class – yes, which equals to TP/(TP+FN)), and F1 score (i.e., the weighted average of precision and recall, which equals to 2 * (Recall * Precision) / (Recall + Precision)).

In our results, the LDA model with the number of clusters pre-defined as 6 could not provide distinguished top words for each ethnicity to perform classification prediction. For example, the group Non-Hispanic American Indian and Alaska Native does not obtain any top words (i.e., surnames) to be identified in a cluster, while the Non-Hispanic White group obtain 4 clusters with significant top words to be considered as Non-Hispanic White ethnicity. Therefore, the original LDA model would not be ideal for the classification of our dataset.

The sLDA model gains a better prediction result than the LDA model. Although some ethnicity groups (e.g., Non-Hispanic Black or African American / Hispanic or Latino origin) are considered as one cluster by the sLDA model, the metrics values show that this model can reach an overall satisfactory result for individual ethnicity prediction, especially for the groups of Non-

Hispanic Asian and Native Hawaiian and Other Pacific Islander and Non-Hispanic American Indian and Alaska Native ethnicity.

As for the Census Model, it outperforms the sLDA model with the higher metrics values as well as the more specific subdivisions of ethnicity groups. Therefore, the results of Census Model will be utilized for subsequent movement pattern analysis. The reason why the Census Model works well might be due to the fact that the ethnicity distribution of the Twitter population is similar to the one of the national population. In 2014, the Pew Research Center reported that among all online adults, 25% of Non-Hispanic Black or African American as well as Hispanic or Latino origin were twitter users along with 21% of Non-Hispanic White. Our result shows that the distribution of Non-Hispanic White, Non-Hispanic Black or African American, Non-Hispanic Asian and Native Hawaiian and Other Pacific Islander, Non-Hispanic American Indian and Alaska Native, Non-Hispanic Two or More Races, Hispanic or Latino origin is 63.50%, 11.05%, 5.06%, 0.79%, 1.31%, 18.29%, compared with the country level 62.2%, 12.4%, 5.4%, 0.7%, 2.0%, 17.4% respectively from the U.S. Census ACS 2010-2014.

4.2. Tweeting Density of Different Groups

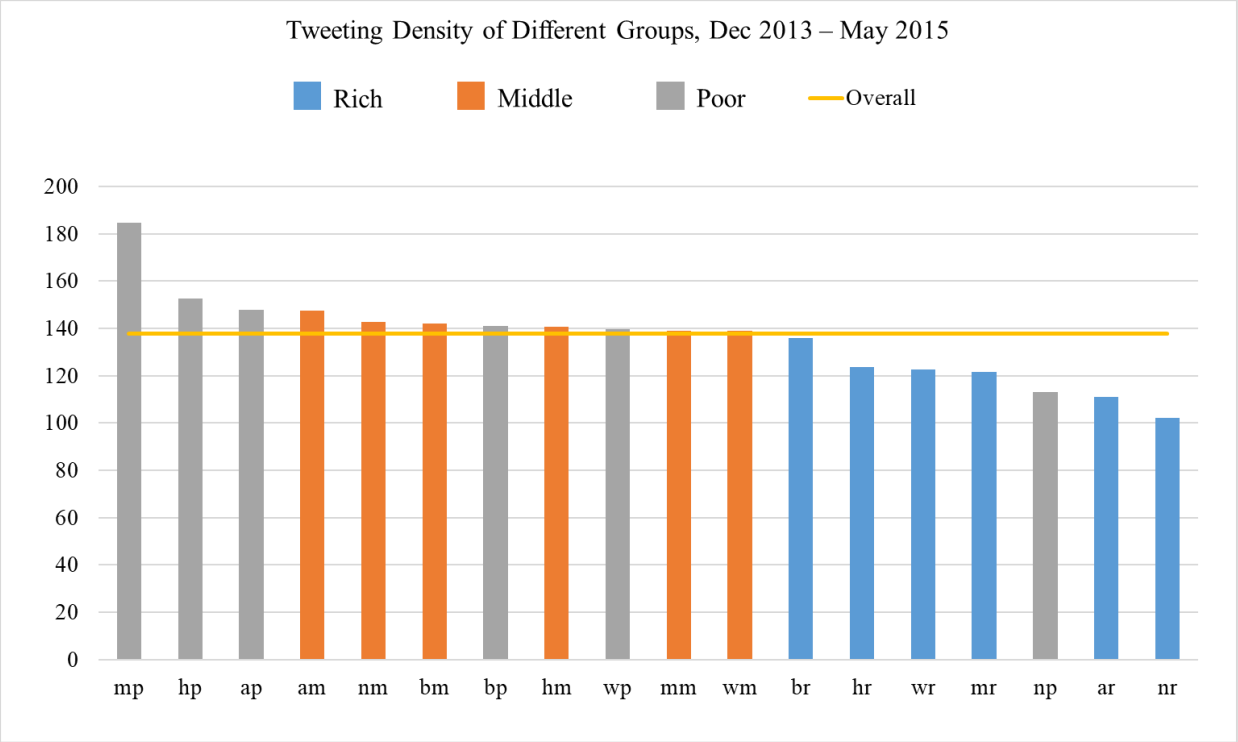


Figure 4. Tweeting density of different groups from December 2013 to May 2015. For the purpose of simpler representation, each group is represented by two characters, the first one stands for race-ethnicity, and the second one for economic status. Specifically, the corresponding relationships are w - Non-Hispanic White, b - Non-Hispanic Black or African American, a - Non-Hispanic Asian and Native Hawaiian and Other Pacific Islander, n - Non-Hispanic American Indian and Alaska Native, m - Non-Hispanic Two or More Races, and h - Hispanic or Latino origin respectively; r - rich, m - middle and p - poor. For example, wr represents the group of Non-Hispanic White and Rich

The tweeting density of one group is defined as the tweeting frequency divided by the population of the group, which indicates the average number of tweets that are posted by each individual from one group within the 18 months. Figure 4 shows that the tweeting density ranges

from 102 (Non-Hispanic American Indian and Alaska Native + Rich) to 185 (Non-Hispanic Two or More Races + Poor), and the overall level (i.e., mean) is 138 with standard deviation as 19. All groups are within the ±2 standard deviations of the mean value except for the group Non-Hispanic Two or More Races + Poor. Therefore, although the tweeting density among different racial-ethnic and economic groups slightly vary from each other, yet we can still conclude that the twitter posts can capture similar amounts of activity zones among different racial-ethnic and economic groups.

To explore further, poor and middle groups share a more similar tweeting density with an average value of 144 and 140 respectively, while rich groups (average 123) have a lower tweeting density (17% lower than poor groups). As for the racial-ethnic factor, the range of average values from different racial-ethnic groups is from 136 to 142, thus there is no much difference among the racial-ethnic groups.

4.3. Movement Pattern Analysis

4.3.1. Average number of activity zones for different groups

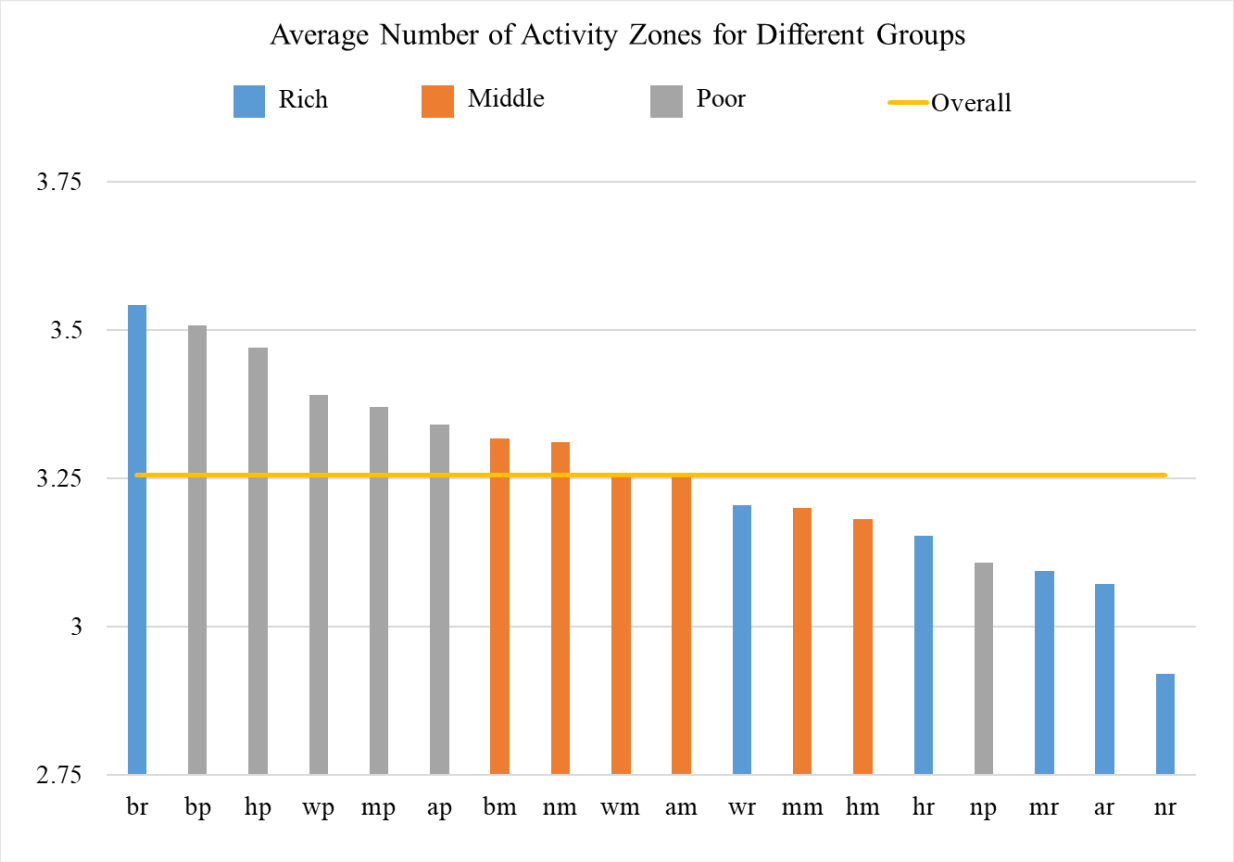


Figure 5. Average number of activity zones for different groups

In general, poor groups have more activity zones than rich groups (Figure 5). The average number of activity zones for poor groups is 3.42, along with middle groups 3.25 and rich groups 3.23. This could be due to the fact that rich groups post less tweets (see Section 4.2) so that their activity zones are less likely to be captured.

However, for the racial-ethnic factor, the average numbers of activity zones range from 3.20 for Non-Hispanic Two or More Races and Hispanic or Latino origin to 3.36 for Non-Hispanic Black or African American. With the same tweeting density (140) of these three racial-ethnic

groups, it indicates that Non-Hispanic Black or African American have 5% more activity zones than Non-Hispanic Two or More Races and Hispanic or Latino origin.

On the other hand, the groups Non-Hispanic Black or African American + Rich and Non-Hispanic Black or African American + Poor, whose tweeting densities are both at the average level (see Section 4.2), have the most average number of activity zones among all groups. Another finding is that within both rich and poor groups, race-ethnicity has more impacts on the variations of average number of activity zones.

4.3.2. Spatial variability and demographic differences of travel distances

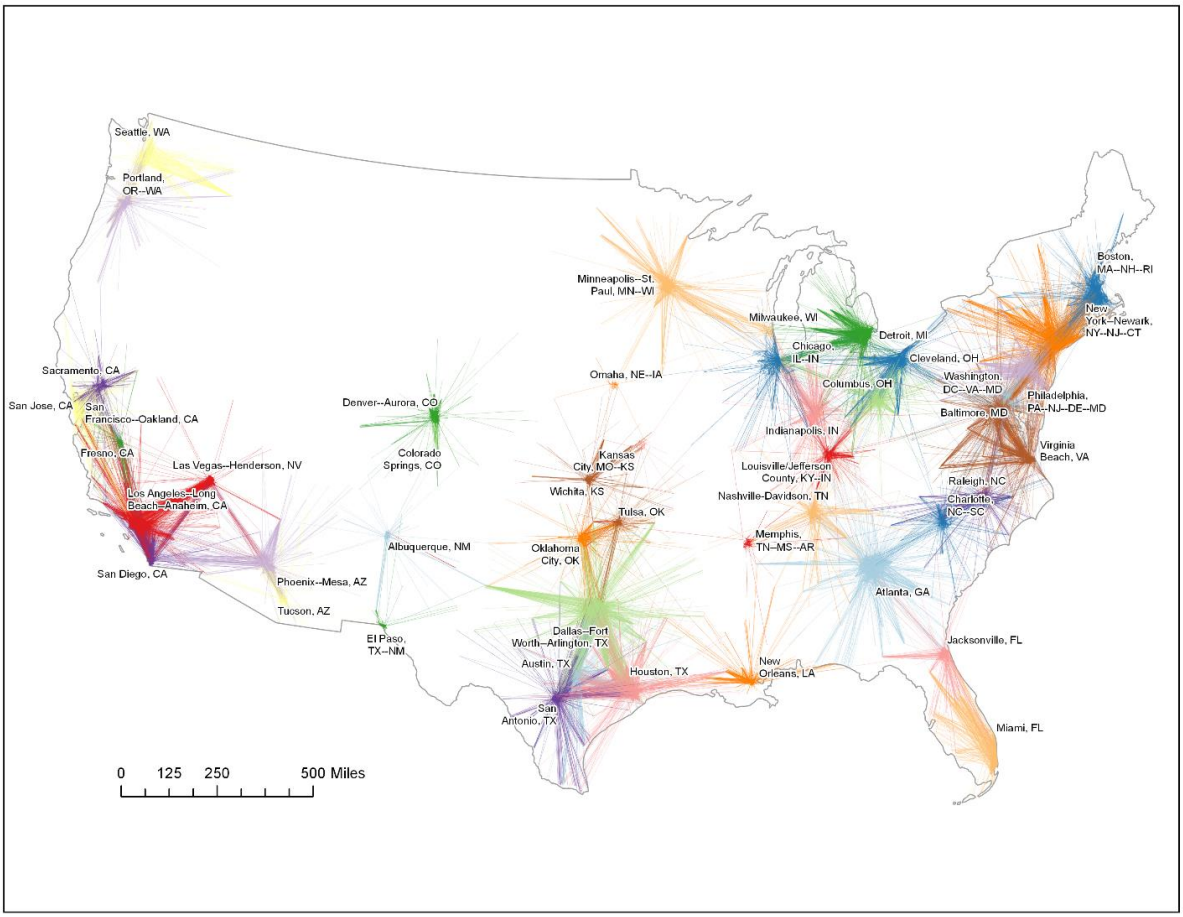


Figure 6. Twitter users' travels from predicted homes to other activity zones among the U.S. top 50 cities (> 500 km travels excluded)

The inter-city interactions can be perceived from Figure 6. With all travels less than 500 km among the U.S. top 50 cities, the spatial variability of the spread destinations can indicate the different interactions among those cities. For example, New York appears strong connections with Boston, Washington DC and Philadelphia, while Los Angeles shows a similar pattern with Las Vegas, San Diego, Long Beach, and the Bay Areas. Smaller, inland cities such as Omaha and Albuquerque are less connected to other cities.

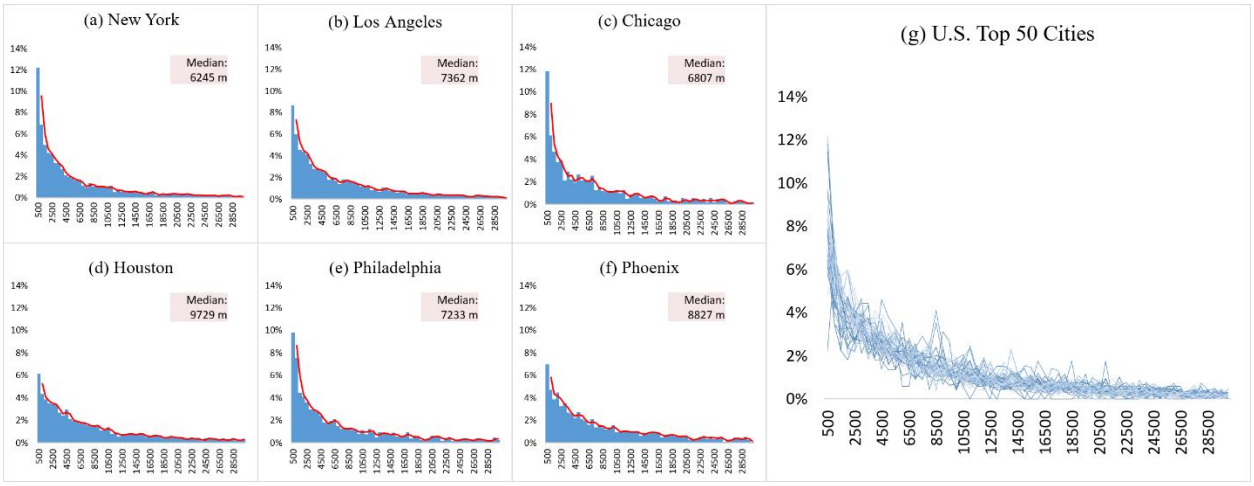


Figure 7. The travel distance distributions for individuals' daily activity zones in the U.S. top 6 cities (a-f) and travel distance distributions of the U.S. top 50 cities (g)

From Figure 7g, we can observe a uniform pattern of travel distance distributions among the U.S. top 50 cities, which are fitting a decreasing curve with long-tail. Particularly, the median travel distance for New York City is 6245 m (Figure 7a); Los Angeles is 7362 m (Figure 7b);

Chicago is 6807 m (Figure 7c); Houston is 9729 m (Figure 7d); Philadelphia is 7233 m (Figure 7e); and Phoenix is 8827 m (Figure 7f). This finding shows that a larger proportion of residents from public-transit-friendly cities, e.g., New York City and Chicago, can travel shorter distances compared with those from car-dependent cities, e.g., Houston. Note that some cities, such as Los Angeles, which have built and enhanced public transportation systems and bike lane networks in the last decade, offer strong alternatives to driving cars — and cut down on the residents' travel distances as well.

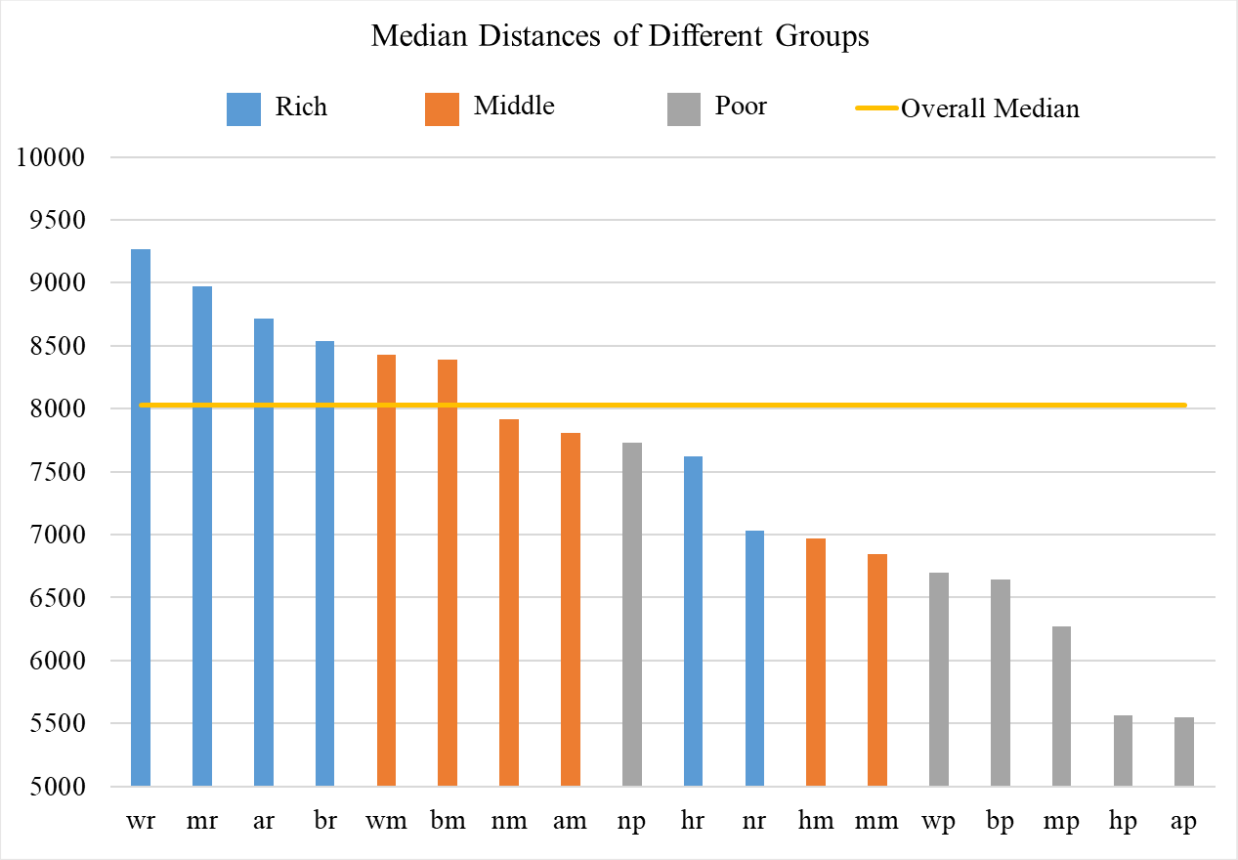


Figure 8. Median travel distances for different groups

From Figure 8, we can obtain more detailed findings of the median travel distances among different racial-ethnic and economic groups. Generally speaking, poor groups travel shorter distances to their activity zones than rich and middle groups, especially for the ones that are race-ethnicity minorities such as Non-Hispanic Asian and Native Hawaiian and Other Pacific Islander as well as Hispanic or Latino origin. Specifically, the median travel distance of poor groups is 6240 m, along with middle groups 8034 m and rich groups 8872 m. That is, poor groups are 42% shorter in median travel distance than rich groups. This finding supports the theories from previous urban and transportation studies - economically disadvantaged population has less disposable income, and they have to rely on public transportation more than the affluent people, their mobility is therefore restricted (Murakami and Jennifer 1997, Giuliano 2005, Paulley, Balcombe et al. 2006). Although the race-ethnicity factor can differ from the theories, such as Non-Hispanic American Indian and Alaska Native + Rich having shorter median travel distances than the Non-Hispanic American Indian and Alaska Native + Poor, yet the abnormality might be due to the insufficient amount of social media users identified in this minor race-ethnicity group.

On the other hand, among different racial-ethnic groups, the median distances are 8450 m, 8304 m, 7760 m, 7682 m, 7153 m, 6868 m for Non-Hispanic White, Non-Hispanic Black or African American, Non-Hispanic American Indian and Alaska Native, Non-Hispanic Asian and Native Hawaiian and Other Pacific Islander, Non-Hispanic Two or More Races, Hispanic or Latino origin respectively. It indicates that the racial-ethnic groups of Non-Hispanic White as well as Non-Hispanic Black or African American travel longer median distances than other racial-ethnic minorities, which might indicate that racial-ethnic minorities have lower social status in relation to their daily use of space and limited integration into society. Particularly, the median travel distance of Non-Hispanic White is 23% longer than the one of Hispanic or Latino origin, 18%

longer than the one of Non-Hispanic Two or More Races, 10% longer than the one of Non-Hispanic Asian and Native Hawaiian and Other Pacific Islander, 9% longer than the one of Non-Hispanic American Indian and Alaska Native.

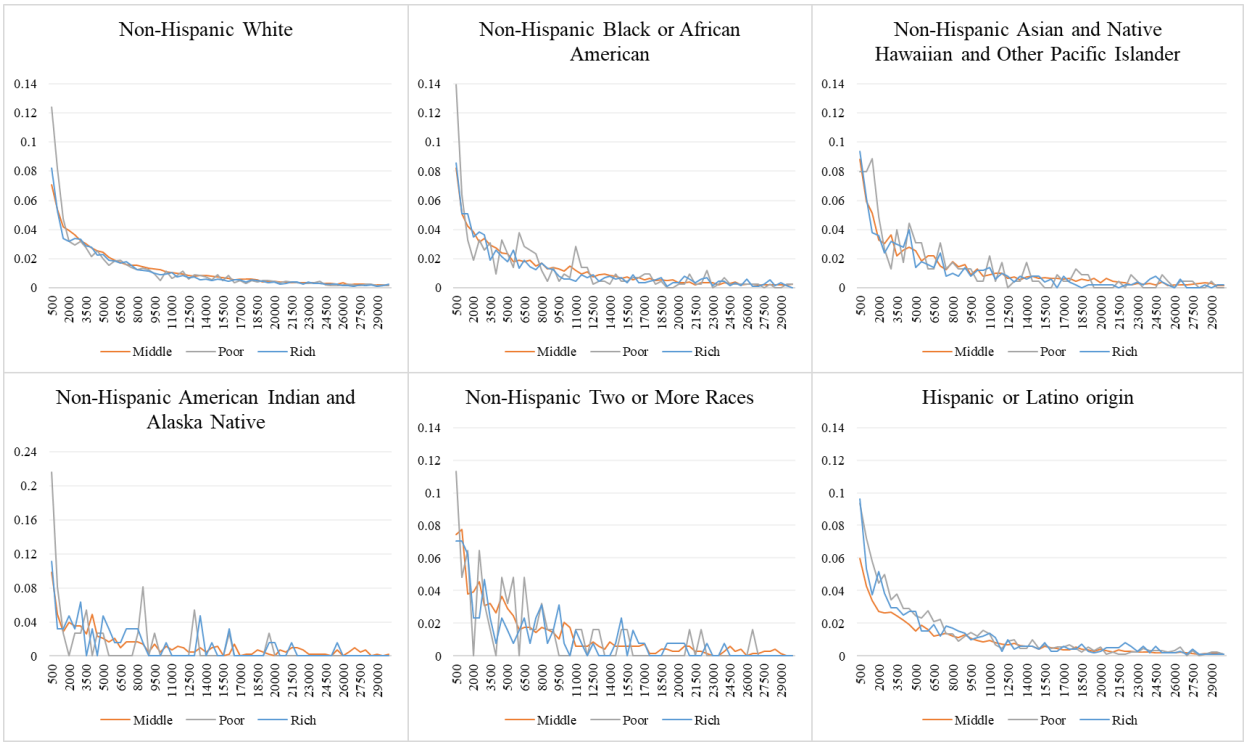


Figure 9. Travel distance distribution of different racial-ethnic and economic groups (< 30 km)

To explore in depth, the travel distance distributions (< 30 km) among different groups are displayed in Figure 9. There are three distinguished patterns: 1) the Non-Hispanic White group has the most stable pattern of the travel distance distribution – less and less people travel if the distance increases; 2) poor groups from the six ethnicities generally have a larger proportion in shorter distance travels (< 0.5 km); and 3) the proportion changes along with the increase of travel distance from poor groups are more fluctuated and erratic than the ones from rich and middle

groups, especially for the race-ethnicity minorities such as Non-Hispanic Asian and Native Hawaiian and Other Pacific Islander, Non-Hispanic American Indian and Alaska Native, and Non-Hispanic Two or More Races. This finding indicates that race-ethnicity and economic status can contribute to significant differences in human movement patterns.

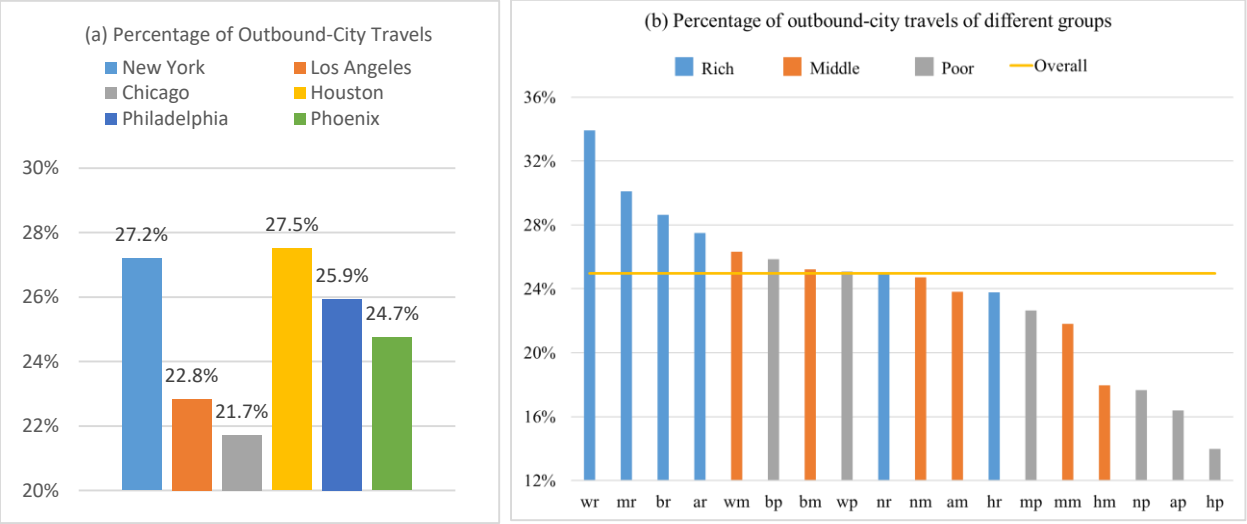
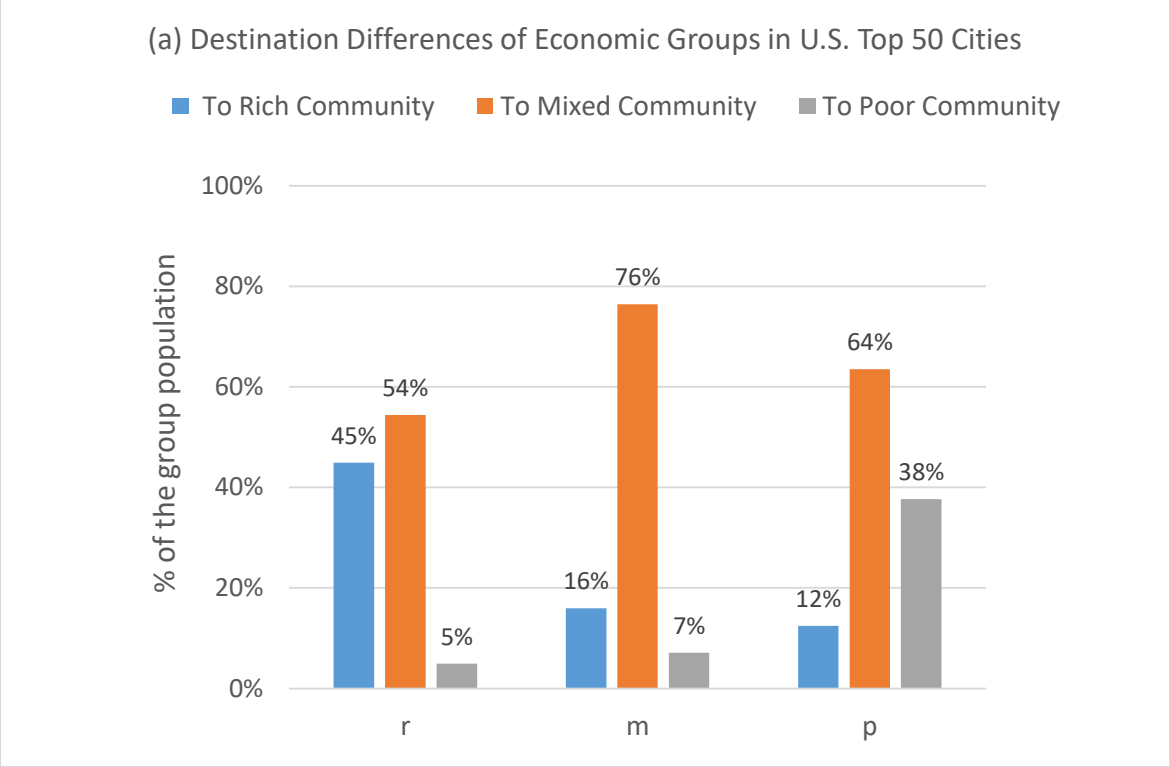


Figure 10. Percentage of outbound-city travels of the U.S. top 6 cities (a), and of different groups from the U.S. top 6 cities (b)

To further explore the mobility differences in longer distance travels, the outbound-city travels from are extracted out to learn the spatial variability and demographic differences. Figure 10a shows the percentage of outbound-city travels of the U.S. top 6 cities. Specifically, New York (27.2%) and Houston (27.5%) have more outbound-city travels, which could indicate their lightly stronger interaction power with other cities, while Los Angeles (22.8%) and Chicago (21.7%) have less outbound-city travels.

Figure 10b displays the percentage of outbound-city travels of different groups from the U.S. top 6 cities. Obviously, poor groups (average 22%) contribute less outbound-city travels than rich groups (average 32%) and middle groups (average 24%), and the overall mean percentage is 24%. Particularly, the poor groups from the racial-ethnic minorities such as Non-Hispanic American Indian and Alaska Native, Non-Hispanic Asian and Native Hawaiian and Other Pacific Islander, and Hispanic or Latino origin have much lower percentages, while the poor groups from Non-Hispanic White and Non-Hispanic Black or African American reach the overall mean percentage. Similar to poor groups, middle groups from Non-Hispanic White and Non-Hispanic Black or African American have the highest percentages of outbound-city travels. In addition, the Hispanic or Latino origin + Middle group shows a relatively low percentage. This finding strongly proves that people who are economically disadvantaged and racial-ethnic minorities are more restricted in long distance travels, which indicates their spatial mobility is more limited into the local scale.

4.3.3. Urban mobility spread in different economic destinations


(a) Destination Differences of Economic Groups in U.S. Top 50 Cities
To Rich Community
To Mixed Community
To Poor Community
% of the group population
100%
80%
60%
40%
20%
0%
45%
54%
5%
16%
76%
7%
12%
64%
38%
r
m
p

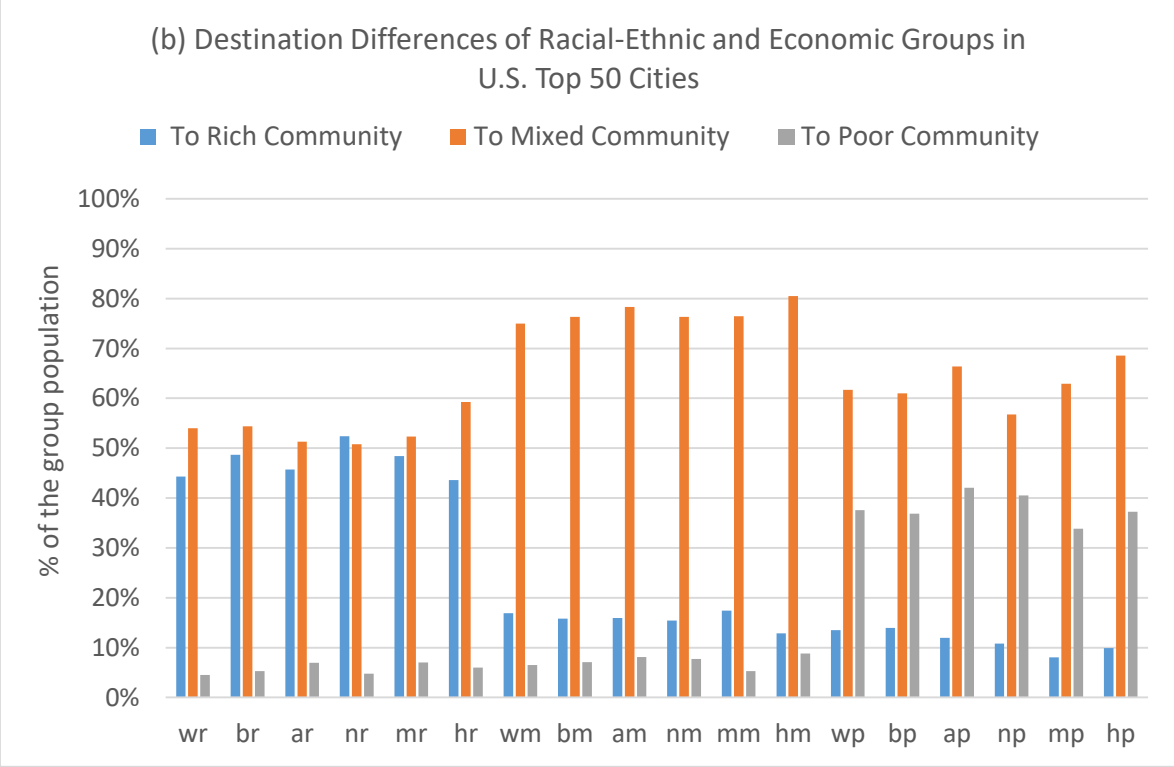

(b) Destination Differences of Racial-Ethnic and Economic Groups in U.S. Top 50 Cities
To Rich Community
To Mixed Community
To Poor Community
% of the group population
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
wr
br
ar
nr
mr
hr
wm
bm
am
nm
mm
hm
wp
bp
ap
np
mp
hp

Figure 11. Percentage of different economic groups traveling to rich, middle, and poor communities in the U.S. top 50 cities (a), and percentage of different racial-ethnic and economic groups traveling to rich, middle, and poor communities in the U.S. top 50 cities (b)

Figure 11a displays an economically-segregated movement pattern among different economic groups in the U.S. top 50 cities - rich neighborhoods are mostly visited by the rich, while poor neighborhoods are mainly accessed by the poor. Specifically, 45% of the rich have access to rich neighborhoods as their activity zones, compared to 16% of the middle and 12% of the poor; 38% of the poor travel to poor neighborhoods as their activity zones, while only 5% of the rich and 7% of the middle visit poor neighborhoods. With race-ethnicity as consideration in Figure 11b, there is no much difference in the national scale.

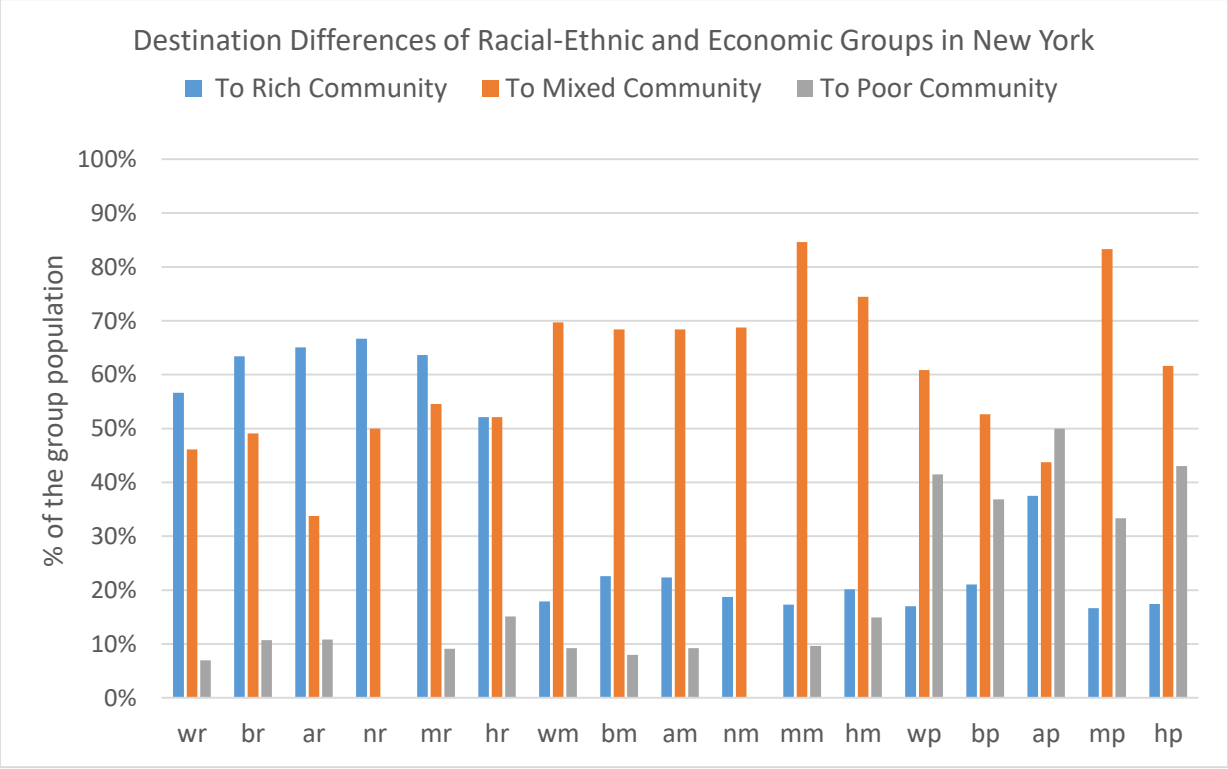


Figure 12. Percentage of different racial-ethnic and economic groups traveling to rich, middle, and poor communities in New York

Figure 12 takes New York City as an example to show the percentage of different racial-ethnic and economic groups traveling to rich, middle, and poor communities. The economically-segregated travel pattern in New York City is overall similar to the national pattern. R-to-r travels range from 52% (Hispanic or Latino origin + Rich) to 67% (Non-Hispanic American Indian and Alaska Native + Rich), and all are higher than the national level (45%). P-to-r travels range from 17% (Non-Hispanic Two or More Races + Poor) to 38% (Non-Hispanic Asian and Native Hawaiian and Other Pacific Islander + Poor), and all are higher than the national level (12%) as well. This finding indicates that the rich community is more likely to be accessed in New York City.

Particularly, among all poor groups, the Non-Hispanic Asian and Native Hawaiian and Other Pacific Islander + Poor group shows the best accessibility to the rich community (38%) as well as to the poor community (50%). When we map out their trajectories to visualize the destinations (Figure 13), we can see that they are more likely to work in Manhattan CBD, and there is a residential cluster in Downtown Flushing. Surprisingly, this discrepancy between home and work locations can refer to the concept of “spatial mismatch” - the mismatch between housing market segregation/discrimination and the employment and earnings of the racial-ethnic minority Non-Hispanic Black or African American (Holzer 1991, Kain 2004). Kain added that serious restrictions on the residential choices of Chicago and Detroit Black residents negatively affected their welfare, such as housing market discrimination on housing prices, home-ownership and educational opportunities. Specifically, the continued Black segregation and the resulting

concentration of Black children in low-achieving schools accounted for a large part of the Black–White achievement gap. Therefore, this “spatial mismatch” phenomenon deserves more attention when it shows up for the Non-Hispanic Asian and Native Hawaiian and Other Pacific Islander + Poor group in New York.

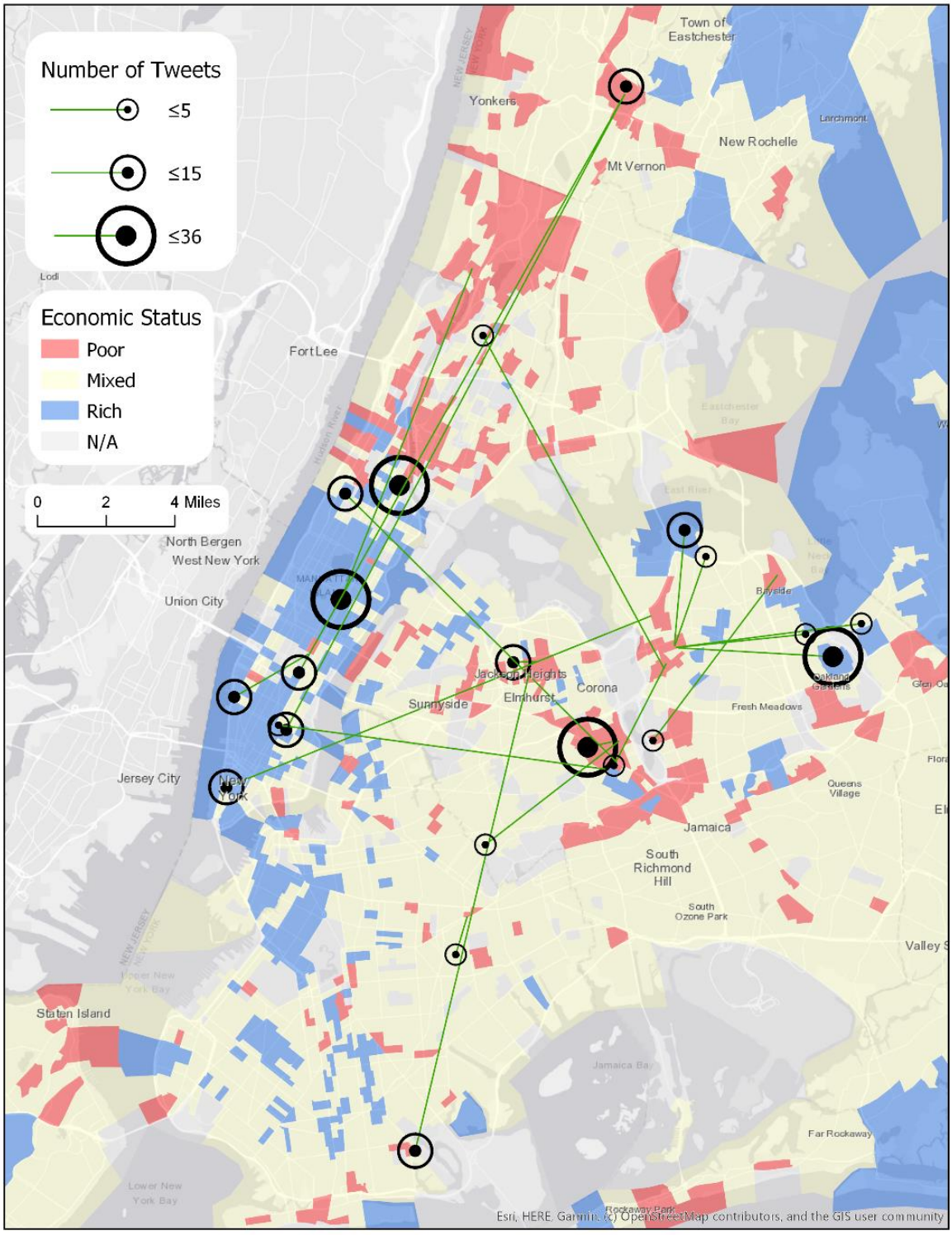


Figure 13. Inner-city travels from home to activity zones of Non-Hispanic Asian and Native Hawaiian and Other Pacific Islander + Poor group in New York

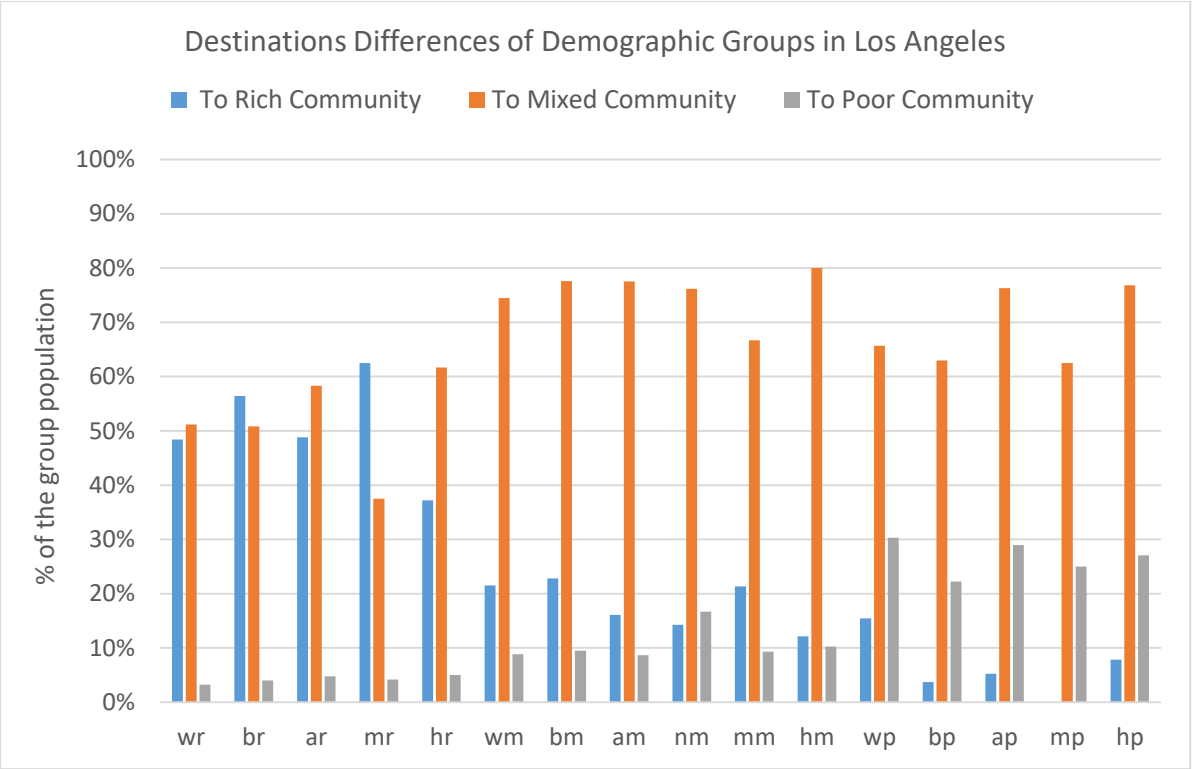


Figure 14. Percentage of different racial-ethnic and economic groups traveling to rich, middle, and poor communities in Los Angeles

Figure 14 takes Los Angeles as an example to show the percentage of different racial-ethnic and economic groups traveling to rich, middle, and poor communities. The segregated travel pattern in Los Angeles is also overall similar to the national pattern.

For r-to-r travels, only the Hispanic or Latino origin + Rich group (37%) is lower than the national level (45%). P-to-p travels range from 22% (Non-Hispanic Black or African American + Poor) to 30% (Non-Hispanic White + Poor), and all are lower than the national level (38%). Meanwhile, for p-to-r travels, only the Non-Hispanic While + Poor group (15%) is higher than the national level (12%), and all other poor groups are much lower than 12%.

The locations of homes (Figure 15a) and activity zones (Figure 15b) of the Non-Hispanic While + Poor group are shown in detail. The majority of home locations are clustered in the poor census tracts in the southern middle areas of Los Angeles as well as in Pomona and Ontario. Clearly, there is an economically-segregated residential pattern in Greater Los Angeles. On the other hand, most activity zones are in the vicinity of home locations, and there are two apparent activity zone clusters in Ontario and Fullerton (the neighborhoods near California State University - Fullerton). Besides these two clusters, activity zones are dispersedly located at the middle census tracts that are close to the home locations, and a few of them are in more affluent neighborhoods such as Beverly Hills, Santa Monica, Huntington Beach, etc., which explains why this group has a higher proportion of going to the rich community.

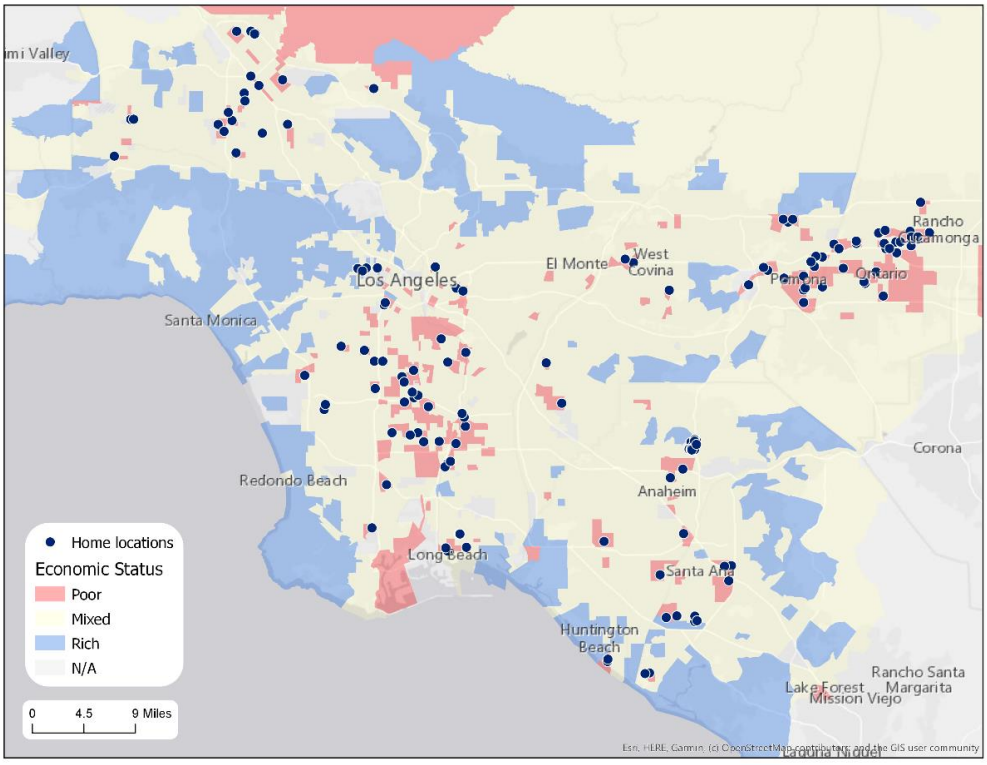


Figure 15a. Home of Non-Hispanic White + Poor group in Los Angeles

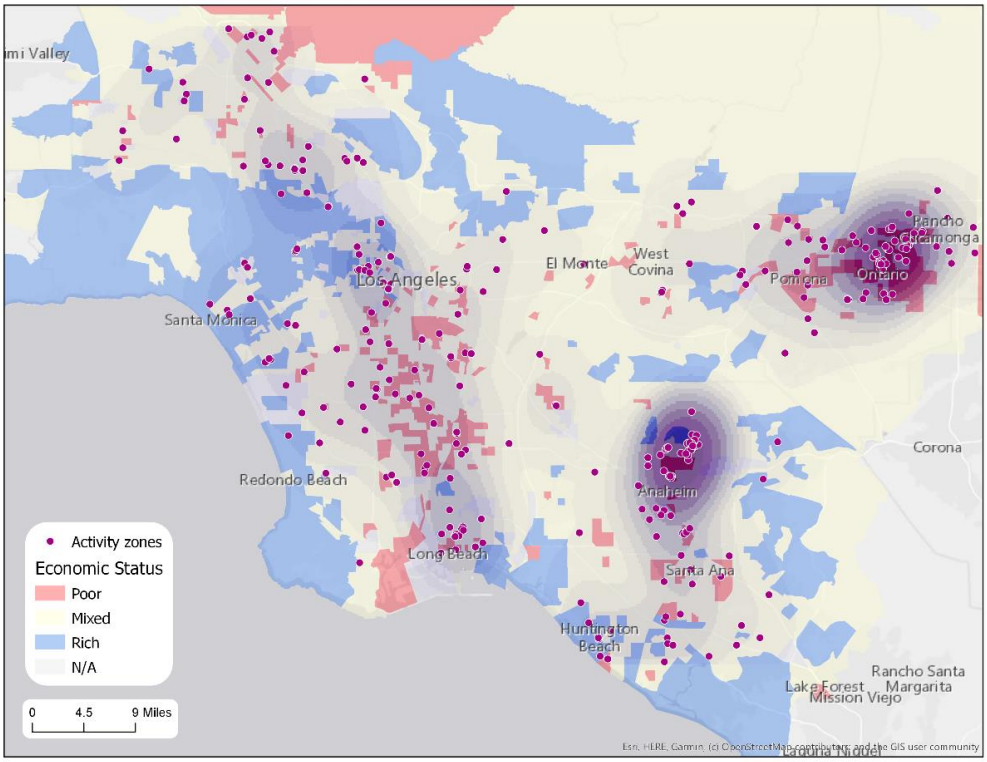


Figure 15b. Activity Zones of Non-Hispanic White + Poor group in Los Angeles

Chapter 5. Conclusion and Discussion

This thesis aims to summarize and compare the movement patterns among different racial-ethnic and economic groups using social media from the U.S. top 50 populated cities. It reveals significant findings as below:

1) The average tweeting density of rich groups is 17% lower than the one from poor groups, while the average tweeting densities among different racial-ethnic groups show similarity.
2) With the same tweeting density, Non-Hispanic Black or African American have 5% more activity zones than Non-Hispanic Two or More Races and Hispanic or Latino origin. Moreover, the groups Non-Hispanic Black or African American + Rich and Non-Hispanic Black or African American + Poor, whose tweeting densities are both at the average level (see Section 4.2), have the most average number of activity zones among all groups.
3) Poor groups are 42% shorter in median travel distance than rich groups. On the other hand, the racial-ethnic groups of Non-Hispanic White as well as Non-Hispanic Black or African American travel longer median distances than other racial-ethnic minorities, which might indicate that racial-ethnic minorities have lower social status in relation to their daily use of space and limited integration into society. Particularly, the median travel distance of Non-Hispanic White is 23% longer than the one of Hispanic or Latino origin, 18% longer than the one of Non-Hispanic Two or More Races, 10% longer than the one of Non-Hispanic Asian and Native Hawaiian and Other Pacific Islander, 9% longer than the one of Non-Hispanic American Indian and Alaska Native.
4) Poor groups contribute 10% less outbound-city travels than rich groups. Particularly, the poor groups from the racial-ethnic minorities such as Non-Hispanic American Indian and Alaska Native (18%), Non-Hispanic Asian and Native Hawaiian and Other Pacific Islander

(16%), and Hispanic or Latino origin (14%) have much lower percentages of outbound-city travels, while the poor groups from Non-Hispanic White and Non-Hispanic Black or African American reach the overall mean percentage (25%). Similarly, middle groups from Non-Hispanic White (26%) and Non-Hispanic Black or African American (25%) reach the highest percentages of outbound-city travels among all middle groups, while the Hispanic or Latino origin (18%) shows the lowest percentage among all middle groups. This finding strongly proves that people who are economically disadvantaged and racial-ethnic minorities are more restricted in long distance travels, which indicates their spatial mobility is more limited into the local scale.

5) An economically-segregated movement pattern in the national scale is observed – rich neighborhoods are mostly visited by the rich, while poor neighborhoods are mainly accessed by the poor, but some race-ethnicity groups can diversify this segregated pattern in the local scale, such as the Non-Hispanic Asian and Native Hawaiian and Other Pacific Islander + Poor group in New York having a much higher percentage (38%) of traveling to rich community than the national average level of p-to-r travels (12%).
6) The spatial variability of travel distances is also revealed. Although there is a uniform pattern of travel distance distributions among the U.S. top 50 populated cities, which are fitting a decreasing curve with long-tail, yet the median travel distance for the top 6 cities are significantly different (e.g., New York City 6245 m; Los Angeles 7362 m; Chicago 6807 m; Houston 9729 m; Philadelphia 7233 m; and Phoenix 8827 m). This finding shows that a larger proportion of residents from public-transit-friendly cities, e.g., New York City and Chicago, can travel shorter distances compared with those from car-dependent cities, e.g., Houston. On the other hand, for the percentage of outbound-city travels of the U.S.

top 6 cities, it shows that New York (27.2%) and Houston (27.5%) have more outbound-city travels, which could indicate their lightly stronger interaction power with other cities, while Los Angeles (22.8%) and Chicago (21.7%) have less outbound-city travels.

In this thesis, we acknowledge that privacy in social media is a sensitive subject when we display and analyze the results, therefore, we always respect and protect privacy at the individual level. Both our analysis of collective movement pattern and the revelation of spatial variability do not expose any individual's private information (e.g., predicted race-ethnicity, inferred home location, inferred economic status). For example, Figure 6 shows Twitter users' travels from predicted homes to other activity zones among the U.S. top 50 cities (> 500 km travels excluded), which displays the spatial variability of travel distances aggregated into urban area level. Similarly, exploring the collective movement patterns will be ideal for protecting privacy since all individuals are represented by different racial-ethnic and economic groups.

Another major significance of this thesis is that our novel methodology for the analysis of collective movement patterns is applicable to social media data (e.g., geotagged tweets) and beyond, such as cell phone location data and travel diary with carry-on GPS devices. Therefore, the methodology workflow will be a significant prototype in future for exploring the collective movement patterns using any point-based data source.

However, due to the source of the data and the methodology to process the data, this research may potentially have several limitations and some of them could be improved in future work as below:

1) Data might not be representative for the older. According to the Pew Research Center about social media update in 2016, younger Americans are more likely than older Americans to

be on Twitter (Greenwood et al. 2016). Approximately 36% of online adults ages 18-29 are on the social network, more than triple the share among online adults ages 65 and older (10%). Therefore, the twitter users in our users' database are more likely to be a younger generation, and the collective mobility patterns that we unfold in this research may not reflect the patterns of the older generation.

2) Name-ethnicity prediction models can be improved by considering the potential impacts from spatial residential segregation related to race-ethnicity, since the local race-ethnicity distribution could be a considerably influential variable for individual race-ethnicity prediction (Chang, Rosenn et al. 2010). For example, Fiscella and Fremont (2006) combined geocoding and surname analysis and was able to show promise for estimating ethnicity. Similarly, Luo, Cao et al. (2016) detected the ethnicity group associated with a surname, and spatially joined the detected home addresses of the twitter users to the associated US census tracts. They adopted a Bayesian method, known as Bayesian Improved Surname Geocoding method (BISG) (Elliott, Morrison et al. 2009), to integrate the ethnic information implied in the surnames and the demographic profiles of the census tracts where the Twitter users reside. However, the validation of these preliminary results related to spatial clustering method is needed.
3) The accuracy of the LDA based race-ethnicity inference model requires self-adjustment. The model result is sensitive to the LDA parameters (priors) and the accuracy of first name and last name extraction. More effort might be needed to select optimal values for these parameters in order to achieve better fitness of the model in an empirical-based method.
4) The semantic knowledge or purpose of each travel (e.g., go to shopping) can be inferred to reveal more meaningful travel patterns at the both individual and collective level. This

work only utilized the residential areas from the NLCD data to infer the home location for each twitter user. In future, we can infer other types of activities (e.g., work, shopping, and eating) for the detected activity zones by using the NLCD data as well as different open-source datasets (e.g., OpenStreetMap), both of which describe highly-developed areas as commercial zones, open areas described as recreation zones (e.g., parks, golf courts, playfields, etc.), as well as transitions between activity zones (Huang and Wong 2016). Therefore, we could integrate more activity types to study movement patterns.

References

Ambekar, A., et al. (2009). Name-ethnicity classification from open sources. Proceedings of the 15th ACM SIGKDD international conference on Knowledge Discovery and Data Mining, ACM.

Andrade, E. L., et al. (2006). Modelling crowd scenes for event detection. Pattern Recognition, 2006. ICPR 2006. 18th International Conference on, IEEE.

Antipova, A., et al. (2011). "Urban land uses, socio-demographic attributes and commuting: A multilevel modeling approach." Applied Geography **31**(3): 1010-1018.

Bagrow, J. P. and Y.-R. Lin (2012). "Mesoscopic structure and social aspects of human mobility." PloS one **7**(5): e37676.

Birant, D. and A. Kut (2007). "ST-DBSCAN: An algorithm for clustering spatial–temporal data." Data & Knowledge Engineering **60**(1): 208-221.

Black, J. and M. Conroy (1977). "Accessibility measures and the social evaluation of urban structure." Environment and Planning A **9**(9): 1013-1031.

Blei, D. M., et al. (2003). "Latent dirichlet allocation." Journal of machine Learning research **3**(Jan): 993-1022.

Blumenberg, E. and K. Shiki (2007). Transportation assimilation: immigrants, race and ethnicity, and mode choice.

Borah, B. and D. Bhattacharyya (2004). An improved sampling-based DBSCAN for large spatial databases. International Conference on Intelligent Sensing and Information Processing, 2004. Proceedings of, IEEE.

Candia, J., et al. (2008). "Uncovering individual and collective human dynamics from mobile phone records." Journal of physics A: mathematical and theoretical **41**(22): 224015.

Chang, J., et al. (2010). "ePluribus: Ethnicity on Social Networks." ICWSM **10**: 18-25.

Chen, J., et al. (2011). "Exploratory data analysis of activity diary data: a space–time GIS approach." Journal of Transport Geography **19**(3): 394-404.

Comenetz, J. (2016). "Frequently occurring surnames in the 2010 Census." United States Census Bureau.

Crane, R. (2007). "Is there a quiet revolution in women's travel? Revisiting the gender gap in commuting." Journal of the American planning association **73**(3): 298-316.

Dong, X., et al. (2006). "Moving from trip-based to activity-based measures of accessibility." Transportation Research Part A: policy and practice **40**(2): 163-180.

Duggan, M. and J. Brenner (2013). The demographics of social media users, 2012, Pew Research Center's Internet & American Life Project Washington, DC.

Elliott, M. N., et al. (2009). "Using the Census Bureau's surname list to improve estimates of race/ethnicity and associated disparities." Health Services and Outcomes Research Methodology **9**(2): 69.

Fiscella, K. and A. M. Fremont (2006). "Use of geocoding and surname analysis to estimate race and ethnicity." Health services research **41**(4p1): 1482-1500.

Giuliano, G. (2005). "Low income, public transit, and mobility." Transportation Research Record **1927**(1): 63-70.

Gonzalez, M. C., et al. (2008). "Understanding individual human mobility patterns." nature **453**(7196): 779.

Hasan, S., et al. (2013). Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. Proceedings of the 2nd ACM SIGKDD international workshop on urban computing, ACM.

He, Y., et al. (2014). "MR-DBSCAN: a scalable MapReduce-based DBSCAN algorithm for heavily skewed data." Frontiers of Computer Science **8**(1): 83-99.

Holzer, H. J. (1991). "The spatial mismatch hypothesis: What has the evidence shown?" Urban Studies **28**(1): 105-122.

Huang, Q., et al. (2014). From where do tweets originate?: a GIS approach for user location inference. Proceedings of the 7th ACM SIGSPATIAL International Workshop on Location-Based Social Networks, ACM.

Huang, Q., et al. (2016). Mining frequent trajectory patterns from online footprints. Proceedings of the 7th ACM SIGSPATIAL International Workshop on GeoStreaming, ACM.

Huang, Q. and D. W. Wong (2016). "Activity patterns, socioeconomic status and urban spatial structure: what can social media data tell us?" International Journal of Geographical Information Science **30**(9): 1873-1898.

Järv, O., et al. (2015). "Ethnic differences in activity spaces as a characteristic of segregation: A study based on mobile phone usage in Tallinn, Estonia." Urban Studies **52**(14): 2680-2698.

Kain, J. F. (2004). "A pioneer's perspective on the spatial mismatch literature." Urban Studies **41**(1): 7-32.

Kwan, M.-P. (2008). "From oral histories to visual narratives: Re-presenting the post-September 11 experiences of the Muslim women in the USA." Social & Cultural Geography **9**(6): 653-669.

Lauderdale, D. S. and B. Kestenbaum (2000). "Asian American ethnic identification by surname." Population Research and Policy Review **19**(3): 283-300.

Leutzbach, W. (1988). Introduction to the theory of traffic flow, Springer.

Longini, I. M., et al. (2005). "Containing pandemic influenza at the source." Science **309**(5737): 1083-1087.

Longley, P. A., et al. (2015). "The geotemporal demographics of Twitter usage." Environment and Planning A **47**(2): 465-484.

Luo, F., et al. (2016). "Explore spatiotemporal and demographic characteristics of human mobility via Twitter: A case study of Chicago." Applied Geography **70**: 11-25.

Marion, B. and M. W. Horner (2007). "Comparison of socioeconomic and demographic profiles of extreme commuters in several US metropolitan statistical areas." Transportation Research Record **2013**(1): 38-45.

Mehran, R., et al. (2009). Abnormal crowd behavior detection using social force model. Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE.

Morstatter, F., et al. (2013). Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. ICWSM.

Murakami, E. and Y. Jennifer (1997). "Daily travel by persons with low income."

Paulley, N., et al. (2006). "The demand for public transport: The effects of fares, quality of service, income and car ownership." Transport policy **13**(4): 295-306.

Peng, C., et al. (2012). "Collective human mobility pattern from taxi trips in urban area." PloS one **7**(4): e34487.

Porteous, I., et al. (2008). Fast collapsed gibbs sampling for latent dirichlet allocation. Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM.

Ramage, D., et al. (2009). Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1, Association for Computational Linguistics.

Shen, Y., et al. (2013). "Investigating commuting flexibility with GPS data and 3D geovisualization: a case study of Beijing, China." Journal of Transport Geography **32**: 1-11.

Simini, F., et al. (2012). "A universal model for mobility and migration patterns." nature **484**(7392): 96.

Song, X., et al. (2014). Prediction of human emergency behavior and their mobility following large-scale disaster. Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM.

Treeratpituk, P. and C. L. Giles (2012). Name-Ethnicity Classification and Ethnicity-Sensitive Name Matching. AAAI.

Tucker, D. (2005). "The cultural-ethnic-language group technique as used in the Dictionary of American Family Names (DAFN)." Onomastica Canadiana **87**(2): 71-84.

Waddell, P. (2002). "UrbanSim: Modeling urban development for land use, transportation, and environmental planning." Journal of the American planning association **68**(3): 297-314.

Wang, Q., et al. (2018). "Urban mobility and neighborhood isolation in America's 50 largest cities." Proceedings of the National Academy of Sciences **115**(30): 7735-7740.