ABSTRACT

Title of Dissertation:	USING SOCIALLY SENSED BIG DATA TO MODEL PATTERNS AND GEOGRAPHIC CONTEXT OF HUMAN ACTIVITIES IN CITIES
	Cheng FU, Doctor of Philosophy, 2018
Dissertation directed by:	Associate Professor, Kathleen Stewart, Department of Geographical Sciences

Abstract: Understanding dynamic interactions between human activities and land-use structure in a city is a key lens to explore the city as a complex system. This dissertation contributes to understanding the complexity of urban dynamics by gaining knowledge of the interactions between human activities and city land-use structures by utilizing free-accessible socially sensed data sources, and building upon recent research trend and technologies in geographical information science, urban study, and computer science. This dissertation addresses three main questions related to human dynamics: 1) how human activities in an urban environment are shaped by socioeconomic status and the intra-city land-use structure, and how in turn, the knowledge of socioeconomic status-activity relationships can contribute to understanding the social landscape of a city; 2) how different types of activities are located in space and time in three U.S. cities and how the spatiotemporal activity patterns in these cities characterize the

activity profile of different neighborhoods in the cities; and 3) how recent socially sensed information on human activities can be integrated with widely-used remotely sensed geographical data to create a novel approach for discovering patterns of land use in cities that are otherwise lacking in up to date land use information. This dissertation models the associations between socioeconomics and mobility in the Washington, D.C. metropolitan area as a case study and applies the learned associations for inferring geographical patterns of socioeconomic status (SES) solely using the socially sensed data. This dissertation also implements a semi-automated workflow to retrieve activity details from socially sensed Twitter data in Washington, D.C., the City of Baltimore, and New York City. The dissertation integrates remotely-sensed imagery and socially sensed data to model the dynamics associated with changing land-use types in the Washington, D.C.-Baltimore metropolitan area over time.

USING SOCIALLY SENSED BIG DATA TO MODEL PATTERNS AND GEOGRAPHIC CONTEXT OF HUMAN ACTIVITIES IN CITIES

by

Cheng Fu

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2018

Advisory Committee:
Dr. Kathleen Stewart, Chair
Dr. Vanessa Frias-Martinez
Dr. Matthew Hansen
Dr. Grant McKenzie
Dr. Colin Phillips

Dedication

To Yawen Fu (傅亚文), Jianxin Cheng (成建新), Caizhen Jin (金彩珍), and Gaofeng Zhang (张高峰) for their love

Acknowledgements

I would like to thank my committee members for their support on research and the preparation for the manuscript. I would also like to thank Dr. Paul Torrens for his help on the early stage of this dissertation. Thanks to Dr. Ruibo Han, who helps to initialize part of the idea in the first study.

Table of Contents

Dedication ii
Acknowledgements iii
Table of Contents iv
List of Tables vi
List of Figures vii
List of Abbreviations x
Chapter 1: Introduction 1
1. Motivation 1
2. Dissertation Structure 5
2.1. Chapter 2: Human Activity and Socioeconomic Status: Knowledge Discovered from Georeferenced Twitter in Washington D.C. Metropolitan 6
2.2. Chapter 3: Identifying Spatiotemporal Urban Activities through Linguistic Signatures 7
2.3. Chapter 4: Integrating Remotely Sensed Imagery and Activity-based Geographical Information to Sense Built-up Land Use Changes in a US Metropolitan Area 9
2.4. Summary on Innovations 10
3. Data Quality and Limitations when Working with Socially Sensed Data 11
4. Overall Contribution of this Dissertation 13
Chapter 2: Human Activity and Socioeconomic Status: Knowledge Discovered from Georeferenced Twitter in Washington D.C. Metropolitan 15
1. Abstract 15
2. Introduction 16
2.1. Human Activity and Socioeconomic Status (SES) 16
2.2. Social Area Mapping 17
2.3. Deriving Activity-based Communities to Map Social Areas 18
3. Study Area, Data and Data Preprocessing 21
4. General Spatial Activity Patterns 23
5. Analyzing SES-activity Relationships 25
5.1. Main Workflow 25
5.2. Derive SES of Census Tracts by Social Area Analysis 26
5.3. Model the Relationship between SES and Human Activity 29
5.4. An Activity-interaction based Data-drive Approach to Infer Regions with Homogeneous SES 33
5.4.1. Structural Characteristics of the Activity-based Tract Network 36
5.4.2. Community Detection 38
5.5. Neighborhoods Comparison 41
6. Conclusions and Future Work 43
Chapter 3: Identifying Spatiotemporal Urban Activities through Linguistic Signatures 45
1. Abstract 45
2. Introduction 45

3. Data 51
4. Methodology 53
4.1. Extract Activity Topics from Georeferenced Tweets 53
4.1.1. The ST-LDA Model 54
4.1.2. Natural Language Processing Pipeline on Data Preprocessing 54
4.1.3. Interpreting Topics 55
4.1.4. Perplexity and Number of Topics 56
4.2. Validate Spatial and Temporal Features of Extracted Activity Topics .. 59
4.2.1. Temporal Profile of Activity Topics 59
4.2.2. Spatial Signatures of Topics 63
4.2.3. Local Activity Topics 63
4.2.4. Mapping Activity Topic Patterns 65
5. Activity Patterns at the Neighborhood Level: Similarity Within and Between Cities 68
5.1. Measuring Similarity 69
5.2. Similarity Matrix Visualization 70
6. A Case Study 71
7. Conclusions & Future Work 78
Chapter 4: : Integrating Remotely Sensed Imagery and Activity-Based Geographic Information to Sense Built-Up Land Use Changes in US Metropolitan Areas 81
1. Abstract 81
2. Introduction 82
3. Study Area and Data 86
4. Methodology 90
4.1. Deriving ISC Objects 90
4.2. Building Physical Signatures for ISC Objects 92
4.3. Linking Tweets to ISC Objects 92
4.4. Preparing Training and Validation Sets 93
4.5. Building Activity Signatures for ISC Objects 93
4.6. Training and Classification 96
5. Results 98
5.1. Model Performances of 10-fold Cross-validation 98
5.2. Model Performance on Validation Sets 100
6. Sprawl of Residential vs Non-Residential Land in the DC-Baltimore Metropolitan Area 102
7. Discussion 105
8. Conclusion 109
Chapter 5: : Conclusions and Future Work 111
1. Conclusions 111
2. Significant results 114
3. Future Work 116
Bibliography 119

List of Tables

Table 2-1 Abbreviation and meaning for the selected socio-demographic variable 27
Table 2-2 Selected principle components and top loading variables. Variables with absolute weight larger than 0.50 are displayed. The sign of weights indicates direction. 28
Table 3-1 Top 5 most similar neighborhoods in different cities for neighborhood "Dupont Circle" in DC with different topic sets and different similarity metrics. Distance value is displaced under the neighborhood name. 74
Table 3-2 MRR and mean NDCG for the neighborhood suggestion that excludes the neighborhoods in the same city. The subscript COS-JSD means using the suggestion from cosine distance-based ranking as benchmark to evaluate the suggestion from JSD based ranking. Vice versa. 78
Table 4-1 Feature groups and their index for the classifier model (ln stands for natural logarithm). 98
Table 4-2 Detailed classification report of a selected cross-validation on features from both physical and activity signatures (accuracy: 0.87, Kappa coefficient: 0.74, AUC: 0.87). 99
Table 4-3 Model performance of 10-fold cross-validation on three signature combinations. 99
Table 4-4 Detailed classification report on the Validation Set A based on the 100 validation ISC objects 100
Table 4-5 Area-adjusted accuracy and error matrix on the 100 validation ISC objects in Validation Set A. The margin of error is based on 1.96 times of standard error of the estimators, which provides 95% confidence. 101
Table 4-6 Detailed classification report on the Validation Set B based on the 100 validation ISC objects 101
Table 4-7 Area-adjusted accuracy and error matrix on the 100 validation ISC objects in Validation Set B. The margin of error is based on 1.96 times of standard error of the estimators, which provides 95% confidence. 102
Table 4-8 Areas of residential and non-residential using the same approach as Figure 4-6. The unit of the values is km^2. The margin of error is based on 1.96 times of standard error of the estimators, which provides 95% confidence. 103

List of Figures

Figure 1-1 Conceptualization of the key topics in this dissertation drawing on a multi-disciplinary perspective.....5
Figure 1-2 Dissertation structure diagram 6
Figure 1-3 Twitter user's demographics (as of p14, Duggan, 2015).....13
Figure 2-1 Residential population and the population density of Washington, D.C. metropolitan area in 2015 based on 2011-2015 5-year ACS. 22
Figure 2-2 Percentage of White population in the study area. 22
Figure 2-3 Density of tweets in the study area. a) by 100-meter grid b) by census tract. Color ramp breaks are based on a head/tail breaks classification (Jiang, 2013). 25
Figure 2-4 Workflow of the study 26
Figure 2-5 The geography of derived latent components. For the Component 1, a higher value indicates lower SES due to the signs of weights displayed in Table 2-2. 29
Figure 2-6 a. Blue: the radius of gyration in individual groups with different SES. Red: the radius of gyrations in tract groups that individuals' radiuses of gyration are aggregated to tracts and the median of aggregated values are used as the representative value of the tract and the tracts are categorize by SES. b. Blue: the entropy of visited tracts in in individual groups with different SES. Red: the entropy of visited tracts in individual groups with different SES. For all cases, SES groups are categorized by equal intervals on values of Component 1. Larger class number indicates lower SES. 33
Figure 2-7 a. The radius of gyration in individual groups with different SES by states. b. The entropy of visited tracts in individual groups with different SES by states.....33
Figure 2-8 Spatial clusters of Component 1. Due to the sign of the component weights, the High-High clusters are the neighborhoods with low SES compared to their neighbor tracts and the Low-Low neighborhoods clusters represent neighborhoods with high SES.....34
Figure 2-9 The frequency of interaction strengths between any pair of tracts in the graph, and the interactions between adjacent tracts. 36
Figure 2-10 Tract betweenness in the network. 37
Figure 2-11 PageRank score of the tracts. IAD is the Dulles International Airport. 38
Figure 2-12 Spatial pattern of communities detected by the first-round community detection (Level 1 communities). 40
Figure 2-13 Spatial pattern of the communities detected by the second-round community detection (Level 2 communities). The first digit of the community ID is corresponding to the community ID in Figure 2-12. 41
Figure 3-1 Word preprocessing 55
Figure 3-2 Word-cloud of sample topics. A larger font size indicates a higher word probability. Note that the word weights have been normalized to allow comparison between topics.....56
Figure 3-3 RPC for different topic numbers 58
Figure 3-4 Histogram of topic counts by tweets categorized as a topic 58
Figure 3-5 Percentage of aggregated Tweet volume by-hour in the three cities. ..60

Figure 3-6 Temporal profile of per-hour percentage for selected topics. IDs are corresponding to the word-clouds in Figure 3-2. (a) is suggested as “Watching live show”; (b) is suggested as ”Work”; (c) is suggested as “Meals”; and (d) is suggested as “Education”. 62
Figure 3-7 Ri, j for topics in each city 64
Figure 3-8 Word-cloud of (a)Topic 2 and (b) Topic 18..... 64
Figure 3-9 The geography of Topic 18 “Political” in DC. Place A: the White House. Place B: the Capitol Hill. (Tweets in water-body and parks are masked out)..... 66
Figure 3-10 The geography of Topic 23 “Meal” in Baltimore City. Place A: the Horseshoe Casino. Place B: a bar area near the O’Donnell Square Park. Place C: a commercial area with bars and restaurants around the intersection at Roland Ave. and W 36th St..... 66
Figure 3-11 The geography of Topic 23 “Meal” in DC. Place A: the commercial area at Farragut Square. Place B: the commercial area around the Dupont Circle. Place C: the commercial area at Georgetown. D: Chinatown in DC..... 67
Figure 3-12 The geography of Topic 26 “Educational” in New York City. Place A: Columbia University. Place B: the North Academic Center of the City University of New York (CUNY). Place C: LaGuardia Community College. Place D: Bishop Kearney High School..... 67
Figure 3-13 MDS of (a) cosine-distance based ALL_TOPIC distribution of neighborhoods, (b) JSD-based ALL_TOPIC distribution of neighborhoods, (c) cosine-distance based COMMON_TOPIC distribution of neighborhoods, and (d) JSD-based COMMON_TOPIC distribution of neighborhoods. 71
Figure 3-14 Cosine distance between Dupont Circle and the other neighborhoods in BC, DC and NYC. Legends in BC and NYC are coordinated to the legend of DC. 73
Figure 4-1 Geography of the study area. ISC-ACM stands for Impervious Surface Cover Annual Change Map. Tweets were collected from October 2014 to April 2015. Red star is Rockville, MD with details discussed in Figure 4-2. 86
Figure 4-2 North of Rockville, MD (marked as star in Figure 4-1) on the ISC-ACM set layers: a. land use recategorized from the 2010 Maryland Land Use Land Cover Map. b. Change Year layer (time of impervious surface increase), c. Change Duration layer (duration of impervious surface increase in terms of year), d. Change Magnitude layer (percentage of impervious surface increase) 88
Figure 4-3 Frequency distribution of ISC objects by area..... 91
Figure 4-4 Two samples of latent topics derived from the tweet set. Font sizes correspond to word weights in probability distribution. It can be interpreted that Topic a is associated with hair cutting activities and Topic b is about dinner..... 96
Figure 4-5 Relative feature importance of the physical signature and activity signature. The feature groups and indexes are the same as Table 4-1. 100
Figure 4-6 Non-residential and residential area developed between 1986 and 2008 in Washington D.C.-Baltimore region by the three sub data sets. The values of the Training Set are the ground truth from land use maps. The values of the other two labeling sets are based on modeling prediction. 103
Figure 4-7 Residential and non-residential area increases by year using the same approach as Figure 4-6. The smoothed curves are based on the average of a three-year moving window..... 104
Figure 4-8 The increased areas of non-residential and residential places by administrative entities and years by using the same approach as Figure 4-6. 105

Figure 4-9 The Friends Community School on the 2010 Maryland Land Use Land Cover Map (left) and on the Google Maps (right). The land parcel that the school locates (marked as the red star in the official land use map) is mislabeled as pasture, although it was converted to school in 2007. 106
Figure 4-10 Detailed ISC object classification result of Application Set A near Bowie, MD. R: residential. NR: non-residential. 108
Figure 4-11 The ratio of tweets with GPS coordinates in the all tweets collected via Twitter Public Streaming API 109

List of Abbreviations

AGI: Ambient Geographical Information

API: Application Program Interface

CDR: Call Detail Record

GIS: Geographical Information Science

GPS: Global Positioning System

ISC: Impervious Surface Change

LDA: Latent Dirichlet Allocation

NLP: Naturel Language Processing

OSM: OpenStreetMap

POI: Point-Of-Interest

SES: Socioeconomic Status

SNS: Social Network Service

ST-LDA: Single Topic Latent Dirichlet Allocation

TF-IDF: Term Frequency Inverse Document Frequency

UTC: Coordinated Universal Time

VGI: Volunteered Geographical Information

Chapter 1: Introduction

1. Motivation

A city is one of the most complex systems represented as a result of socio-economic drive and planning during the development of human society. Even though many cities in developed countries may be slowing down with respect to massive infrastructure construction, numerous questions remain where an understanding of the dynamics of a city can make a significant contribution, for example, with respect to optimizing urban traffic systems, building sustainable cities, keeping neighborhoods safe and resilient, etc. At the same time, urbanization in developing countries is happening at a rapid pace that requires knowledge on how to plan efficient infrastructures. All of these tasks call for insights into a city's dynamics (i.e., traffic and land use changes) both at a higher, system-level perspective of the diversity of physical and socio-economic processes that rule its residents' daily lives, as well as at a lower or more detailed perspective of how individual and collective habits and decisions shape and impact a city's dynamics.

This dissertation consists of three studies to examine the dynamics of different activities (e.g., moving, shopping, and working) and the associations between these activities and residents' socioeconomic status as well as the layout of land use in cities.

Related research topics have prompted numerous studies in Geography. One such area of study is embedded in location theory that can be traced back to Christaller's Central Place Theory in the 1930s (Christaller, 1933) where a city system is modeled

as a hierarchy with a hexagonal spatial layout based on service capacity. This theory was further expanded by Losch and integrated into Isard's general theory on location (Isard & Smith, 1969), and Alonso's Bid Rent Theory (Alonso, 1964) that suggests a concentric intra-urban land-use structure. In the 1970s, behavioral geographers analyzed the impact of location on individual behaviors, starting from modeling the impact of a city location distribution on consumer behavior using computer simulation (Clark et al., 1970). This research inspired more contemporary economists such as Berry, McFadden, and Krugman to introduce space into economic reasoning and explain how a city system forms, and how populations disperse in space over time.

However, meso-level questions remain about how individuals and collective behaviors are shaped by cities, and how a city's infrastructure, such as the layout of land use and transportation systems, are influenced by human activities. In the early 21st century, work on the physical statistics of non-Brownian motion by Barabasi and Gonzalez renewed an interest in these long-standing questions about human dynamics and showed that we may approach these problems by combining investigations of empirical georeferenced Big Data and complexity theory. This body of work attracted the attention of researchers who have studied spatial complexity as contributing to a new science of cities, including work by Batty, Portugali, Pumain, West, and others.

Over the past 10 years, the prevalence of GPS-embedded devices, e.g., GPS navigators and smart phones, as well as location-based services, such as location-based social media services (SNSs), has made it possible for researchers to access data on individuals' behaviors in space, and model spatial behavior patterns individually or

collectively using data-driven methodologies. Such studies commonly involve data that are contributed voluntarily by users, e.g., spatial data presented as maps on OpenStreetMap (OSM, Haklay, 2010; Zook, et al., 2010), often characterized as *volunteered geographical information* (VGI, Goodchild, 2007), or collected as side products, referred to as *ambient geographical information* (AGI, Stefanidis, Crooks, & Radzikowski, 2011). These data include call detail records (CDRs) from mobile phone carriers (González, Hidalgo, & Barabási, 2008; Toole, et al., 2012; Pei et al., 2014), taxi trajectories (Guo, et al., 2012; Liu, et al., 2012; Yuan, Zheng, & Xie, 2012; Pan, et al., 2013), wireless data service records (Nishi, Tsubouchi, & Shimosaka, 2014a, 2014b), and georeferenced SNS records, e.g., Foursquare (Cranshaw, et al., 2012; Goers, 2013; Saker & Evans, 2016; Zhou & Zhang, 2016) and Twitter (Frias-Martinez, Soguero, & Frias-Martinez, 2012; Lee & Sumiya, 2010; Wakamiya, Lee, & Sumiya, 2011; Hong, et al., 2017). This data-driven research paradigm has been recently conceptualized as *social sensing* (Liu et al., 2015), which is an analog to the well-known remote sensing. These data sets are thus increasingly referred to as *socially sensed data*.

This dissertation makes a significant contribution to increasing our understanding about collective human dynamics in an urban context in order to gain knowledge on the drive of human behavior and support urban planning practices. In a city, human dynamics are closely intertwined with land-use layout and individuals' socioeconomic status (SES) in space, forming a complex system involving spatial, temporal, behavioral, physical and social factors (Figure 1-1). To analyze the geospatial

patterns of human dynamics and the drivers that shape these patterns, this dissertation addresses three key topics:

1. How human activities in an urban environment are shaped by SES and the intra-city land-use structure, and how in turn, the knowledge of SES-activity relationships can contribute to understanding the social landscape of a city.
2. How different types of activities are located in space and time in three U.S. cities and how the spatiotemporal activity patterns in these cities characterize the activity profile of different neighborhoods in the cities.
3. How recent socially sensed information on human activities can be integrated with widely-used remotely sensed geographical data to create a novel approach for discovering patterns of land use in cities that are otherwise lacking in up to date land use information.

Due to the lack of Big Data on human activities across space, we still need more insights on the details of human activities and movements in cities, the relationships between the activities and social and physical drives, e.g., the residents' SES and the land use patterns. The first of these key topics looks at the relationship between SES and activity where land-use is embedded as a deterministic yet latent factor. The second and third topics mainly focus on the relationships between detailed land use and activities, and where SES are involved as an implicit factor that shapes the spatial and temporal patterns of different activities. Three research studies are thus conducted based on each of these key topics (KT).

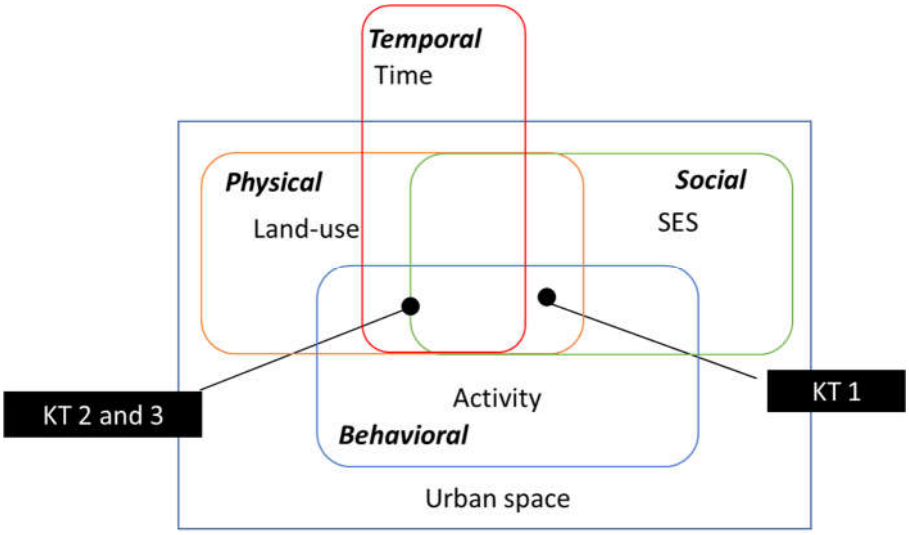


Figure 1-1 Conceptualization of the key topics in this dissertation drawing on a multi-disciplinary perspective

2. *Dissertation Structure*

This dissertation consists of three studies that address the three key topics (Figure 1-2). The overall research goal is to understand the interactions between urban land-use structure and human activity. Chapter 1 introduces the background and overall organization of this dissertation. Both Chapters 2 and 3 model the spatiotemporal patterns of human activities in a city following a data-driven paradigm. Chapter 2 mainly focuses on modeling the associations between general human mobility and residents' SES. It further explores how knowledge of activity-SES associations in turn can help to sense the spatial patterns of SES at the census tract level. Chapter 3 employs natural language processing (NLP) technology to differentiate activity types in three cities to look into the detailed influence of different land uses on different activities. Chapter 4 proposes a methodology for addressing the lack of empirical land-use structure information by applying the activity-land-use models, targeting the goal to fill

the gaps in land use data. Chapter 5 concludes the main findings and innovation of this dissertation and proposes future work.

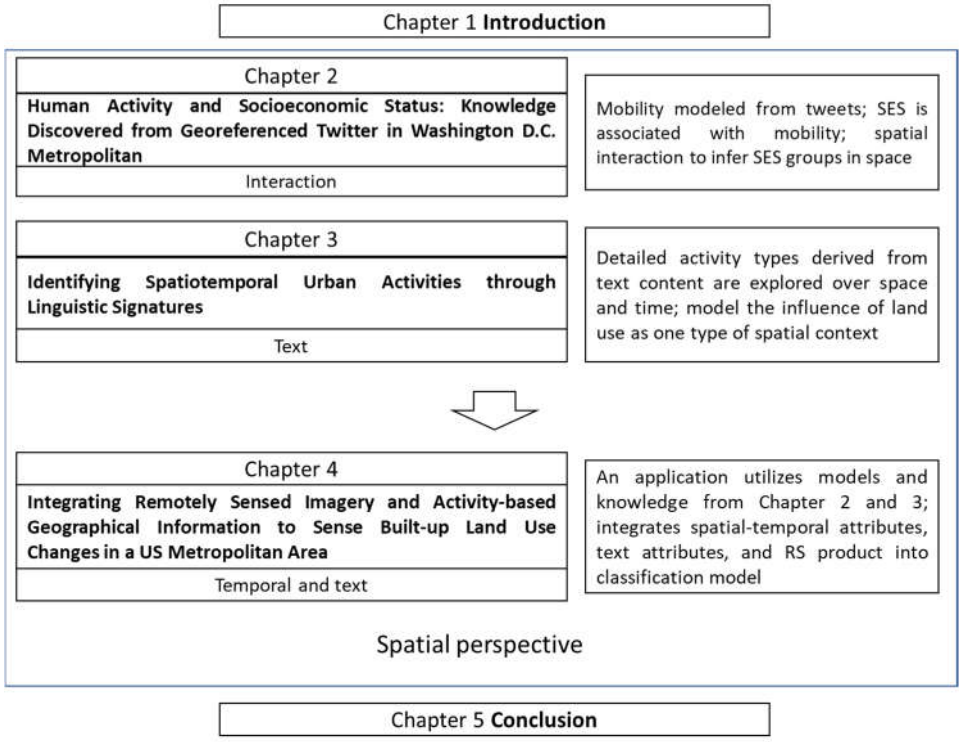


Figure 1-2 Dissertation structure diagram

2.1. Chapter 2: Human Activity and Socioeconomic Status: Knowledge Discovered from Georeferenced Twitter in Washington D.C. Metropolitan

Chapter 2 focuses on investigating the associations between urban residents' SES and their mobility, and applies the results to map the geography of SES by employing a data-driven framework that utilizes socially sensed human activity data. The motivation of this study is to develop a solution for surveying the geographical pattern of SES in cities without regular survey data such as census. Conventional approaches to investigate this topic have relied on travel demand surveys in transportation studies. Regarding its relatively high spatial and temporal resolution as well as high usage penetration for populations all over the world, socially sensed data,

such as Twitter and CDRs, also have the potential to infer the approximate landscape of SES in areas where census and other large socioeconomic surveys are not conducted regularly. This occurs by applying knowledge based on the association between SES and general activity patterns (Longley & Adnan, 2016) and, as this dissertation investigates, the SES-mobility association with sensed activity data.

To achieve the goal, the associations between SES and human mobility in a well-surveyed city, Washington, D.C., are investigated. Community detection in network analysis is further employed to model the inter-tract mobility pattern to discover tract groups that have cohesive intra-group mobility connections. The learned associations between SES and mobility are then applied to infer the relative SES of the tract groups. The inferred SES of tract groups is shown to have good agreement with the census-based SES landscape. This approach sheds light on how social sensing can be applied for mapping the geographical patterns of SES. This study also produces new insights on the complexity of human mobility constrained by SES, physical geography, transportation, and the layout of urban land use.

2.2. Chapter 3: Identifying Spatiotemporal Urban Activities through Linguistic Signatures

The first research topic addresses human activities in general, without distinguishing different activity types and investigates how overall mobility patterns can be shaped by SES and the layout of the city. However, we are in fact often interested in how different activity types are distributed in a city, including the spatio-temporal

details of these activities and encompassing an entire city, providing a lens that reveals the details of urban dynamics.

Chapter 3 looks into different activities that individuals conduct in a city. This chapter presents an approach for modeling the spatiotemporal patterns of different activity types within cities by employing user-contributed, geosocial content as a proxy for human activities. In this work, a semi-automatic workflow mainly relying on topic-based linguistic modeling (Hong et al., 2016) is used to analyze georeferenced twitter data in order to differentiate different activity types. Each extracted topic is a probabilistic distribution of words, whose weights represent the theme of a certain activity semantically, such as shopping, dining, studying, etc. The spatial and temporal patterns of the derived activity types in three U.S. cities: Baltimore, MD., Washington, D.C., and New York City, NY are further examined. The patterns can reflect the linguistic meaning of the activities. This study then constructs a method to link what people post online to the activities conducted within a city.

This study further explores how different neighborhoods in a city are *not* associated with all types of activities in the same way. A novel approach is implemented to characterize city neighborhoods based on the derived set of activities. Each neighborhood is profiled by activity distributions as unique signatures. This research demonstrates how the similarities and differences between neighborhoods can be measured by comparing activity signatures. This further provides an activity-signature-based perspective to describe neighborhoods, which is different from conventional

demographic-signature-based neighborhood profiles, e.g., ESRI Tapestry[1] that mainly reflects the characteristics of *residential* populations based on demography, occupation and income information from the U.S. census.

2.3. Chapter 4: Integrating Remotely Sensed Imagery and Activity-based Geographical Information to Sense Built-up Land Use Changes in a US Metropolitan Area

The results from the analyses undertaken for Chapters 2 and 3 provide insights on how urban land use can lead to heterogeneity of activity distributions in space and time. Chapter 4 applies such knowledge in a practical study on mapping land use structure, i.e., the spatial pattern of non-residential and residential areas, a key component for understanding the complexity of urban systems.

Conventional prevalent land use mapping methods use remotely-sensed imagery-based mapping technology, i.e., remote sensing, and ground surveys provided by government agencies. The major limitation of remote sensing, however, is that sensed imagery can only provide the physical properties of the surface (Herold et al., 2005). Ground surveys are accurate but costly in terms of finances and time, therefore up-to-date official land-use maps by governments are often not widely available for many U.S. cities. Another factor weakening the application of using solely remote-sensing sources for deriving land use maps, is that usage is closely related to human activities as land use is the result of human interaction on the land (Brown, Pijanowski,

[1] http://www.esri.com/landing-pages/tapestry

& Duh, 2000) and land use is less likely to be equivalent to the physical land cover in a post-modern and information-driven economy (Brown, Carolina, & Hill, 2012).

This chapter integrates remotely sensed imagery and socially sensed activity data to infer land use in a metropolitan area. The approach integrates an impervious surface cover change product from remote sensing as the physical signature of land use, and activity signatures derived from georeferenced tweets to infer land use that involves conversions from undeveloped land. A case study is conducted to profile land use change in the Washington D.C.-Baltimore metropolitan area between 1986 and 2008. A classification model utilizing both groups of signatures is developed to differentiate residential and non-residential places. Model assessment shows that the proposed classification workflow can differentiate residential and non-residential uses at an accuracy of over 80%. Combining the temporal information from remotely sensed imagery, the study also reconstructs the temporal trajectory of development for different land use types. Results indicate that the proposed approach is useful for mapping detailed land use in an urban region and serves as a viable new way forward for massive land use surveying that can be more frequent and regular.

2.4. Summary on Innovations

As a multi-disciplinary research effort, this dissertation contributes innovative methodologies and knowledge to both geographical information science and urban geography. For geographical information science, the first study (Chapter 2) of this dissertation shows that integrating network analysis and the complex association between human movement and SES to a human movement model can approximate SES

landscape in a metropolitan area. The second study (Chapter 3) demonstrates a data-driven workflow to retrieve activity details over a city with NLP. The last study (Chapter 4) applies this newly acquired understanding to develop a pipeline for mapping urban land-use structure combining widely accessible remotely sensed data with socially sensed data. With such a pipeline, we may better sense and monitor land use and land-use change in cities and gain more ground truth about urban dynamics from a land-use perspective.

This dissertation also contributes to urban studies in two ways. First, it provides new insights about the complex associations between human movement and SES where no universal association is observed as presented in the first study. This dissertation also shows how intra-city human activities locate in three major U.S. cities, both in terms of overall activity patterns as well as more specific activity types (e.g. working and dining). With such knowledge, we can have a better understanding of the mechanism of urban dynamics as a whole. Finally, the research presented in Chapter 4 observes that non-residential urban land use surpasses residential urban land use after 1996 and that this finding would benefit from further investigation on the drivers for such change.

3. Data Quality and Limitations when Working with Socially Sensed Data

Due to its massive user group, big size, free-to-use policy, and relatively high spatial accuracy, socially sensed data, and particularly the referenced tweets that are used in this dissertation, can be used as a good proxy of human activity. However, the concern about representativeness is also clear. According to a Pew report based on

surveying 1,907 adults, only 20% have accounts (Duggan, 2015). More detailed demographics are displayed in Figure 1-3. In terms of income, there does not appear to be any serious bias. In terms of age, the percentage of younger adults between 18-29 is a little higher than others. Thus, observations from georeferenced tweets may reflect more on young people with activities that are potentially different from other groups. It is still hard to reconstruct a high-resolution individual trajectory solely from georeferenced tweets, although there has been some attempt to do this, but the detected rate is low (Gabrielli et al., 2014). Due to the abnormality of registration, it is also quite hard to validate such results and privacy regulations might be violated. Therefore, it is better to study the dataset from an aggregated perspective. Details on data quality, limitations and potential solutions associated with particular studies are discussed in their corresponding chapters.

Twitter Demographics

Among internet users, the % who use Twitter

	Internet users
Total	*23%*
Men	25
Women	21
White, Non-Hispanic	20
Black, Non-Hispanic (n=85)	28
Hispanic	28
18-29	32
30-49	29
50-64	13
65+	6
High school grad or less	19
Some college	23
College+	27
Less than $30,000/yr	21
$30,000-$49,999	19
$50,000-$74,999	25
$75,000+	26
Urban	30
Suburban	21
Rural	15

Source: Pew Research Center, March 17-April 12, 2015.

PEW RESEARCH CENTER

Figure 1-3 Twitter user's demographics (as of p14, Duggan, 2015)

4. Overall Contribution of this Dissertation

The contribution of this dissertation is two-fold based on understanding how intra-city land-use can impact activity distributions. First, the research for this dissertation models heterogeneity of activity distributions, both by modeling overall activities as well as detailed activity types in urban areas at a macro scale using georeferenced tweets as a proxy for human activities. A framework is proposed to map the SES landscape based on analyzing residents' movements in urban spaces combined with the learned associations between SES and mobility at the socioeconomic group level. This dissertation also demonstrates the usefulness of socially sensed data for determining activity-based neighborhood profiles based on derived activities extracted from tweets to provide a measure for similarity of urban neighborhoods. The results of this research provide new insights about the characteristics of neighborhoods as well as

ways that we can harness this understanding to, for example, find other similar neighborhoods.

Second, the research for this dissertation models how the activity distribution of individual parcels is influenced by different land-use types modeled as parcels at the micro-level, based on activity type and volume over time. Land parcels are derived from remotely sensed imagery-based land cover products. We use the activity models and apply supervised classification models for building an automated mapping workflow to classify derived land parcels based on the activity data. The automated workflows contribute to mapping intra-urban land-use structure in cities for understanding the geography of land-use structure in cities. By combining the temporal information in the remote sensing product, the output of the model can also help to understand the land use change as a procedure.

Chapter 2: Human Activity and Socioeconomic Status: Knowledge Discovered from Georeferenced Twitter in Washington D.C. Metropolitan

1. *Abstract*

Increasingly, knowledge about the influence of socioeconomic status on detailed human activity over space is gained from socially sensed human activities. However, the exact nature of the association between the socioeconomic status and human activity is still an open question on. Using social area as a proxy for socioeconomic status and georeferenced tweets as a proxy for human activities, we propose an analytic framework for determining the association between socioeconomic status and human mobility. For this research, this framework is applied to the city region of Washington, D.C. We find that for this geographic area, the associations between socioeconomic status and human mobility are not universal over the geography, and that the mobility of people with the same socioeconomic status can be influenced by their living location. We apply a data-driven approach to model the activity interactions between census tracts to find tract groups with high activity coherence. This analysis shows that these tract groups spatially co-occur with social area groups that share similar socioeconomic status. However, physical geography is still an important factor to shape mobility patterns even with the well-constructed transportation infrastructure system in the Washington, D.C. metropolitan area. This comprehensive study suggests that the relationship between socioeconomic status and human mobility can vary over space due to location *and* physical geography.

2. *Introduction*

2.1. Human Activity and Socioeconomic Status (SES)

Understanding human activity in cities (e.g., the aggregation and diffusion of people caused by their travel movements) is understood to be an important contributing factor to, for example, the spread of infectious diseases or even epidemics (Merler & Ajelli, 2010; Dalziel et al., 2013), while ongoing activities by residents on urban streets are a possible reason for increased safety in a city (Jacobs, 1961), and intra-city trips by commuters lead to increased demand on transportation infrastructure (Maat, 2009).

Modeling spatiotemporal patterns of human activity including the daily movements of individuals has driven many studies in the past decades. An individual's daily environment for his or her movement is widely conceptualized as *activity space* (Gollege & Stimson, 1997) that is anchored at home and workplace and bounded by other *third places* (Oldenburg & Brissett, 1982). Conventional studies employed sample surveys that used travel diaries that capture individuals' activity spaces. Over the past decade, socially sensed data associated with georeferenced activities (Liu et al., 2015) derived from GPS-embedded devices are widely used for modeling the uneven distribution of human activities in space and time, such as taxi trajectories (Guo, et al., 2012; Liu, et al., 2012), call detailed records (Ratti, et al., 2006; González, Hidalgo, & Barabási, 2008; Reades, Calabrese, & Ratti, 2009; Bajardi, et al., 2015), and georeferenced records on social media, i.e. georeferenced tweets and Foursquare check-ins (Cheng, et al., 2011; Wu, et al., 2014; Jiang, et al., 2016), even though these socially

sensed data have potential bias in representing the social or behavioral profiles of the population.

Certain insights into the drivers of different human mobility patterns both at individual and collective level have been reported. Conventional travel diary surveys have addressed demographic factors, such as gender (Kwan, 1999) and age (Alsnih & Hensher, 2003) on human mobility. SES plays an important role for human activities in general. People with same the same social class tend to live closer and travel longer in daily life (Huang & Wong, 2016; Leo, et al., 2016) and engage in more diverse activities (Pappalardo, et al, 2015). Studies using call data records (CDRs) show that it is also possible to classify individuals' SES given the pattern of phone calls (Smith-Clarke, Mashhadi, & Capra, 2014) or through the application of machine learning approaches on spatial trajectories and statistics on mobility, e.g. travel distance, derived from CDRs (Victor Soto et al., 2011).

2.2. Social Area Mapping

Besides daily activities, SES also shapes urban structure in the form of the residential population distribution. In the fields of urban studies and urban planning, research has been undertaken to understand the influence of SES on urban structure using an analysis framework based on mapping *social area,* which is conceptualized as a group of geographical units, typically census tracts, that shares similar social factors. This analysis framework started by Shevky & Bell (1955) characterizes neighborhoods by three latent dimensions: social rank, economic status, and neighborhood segregation derived from seven census variables. People residing in a social area are conceptualized

as having the same level of living and lifestyle. Even though this analysis framework has been criticized for lacking theoretical support on why the social areas are homogeneous (Hawley & Duncan, 1957), the framework is still in current use and many mathematical tools have been employed to derive the latent social dimensions in a quantitative matter, such as factor analysis (Van Arsdol, Camilleri, & Schmid, 1958; Janson, 1980; Hale & Austin, 1997; Heye, Leuthold, & Bourdieu, 2005), principle component analysis (PCA, Liu & Cao, 2017) and self-organizing maps (Spielman & Thill, 2008). The geographical units in the same social area are not necessary to cluster in space. However, to derive spatially cohesive neighborhoods, the basic geographical units can be further clustered into areas based on their similarity on all or selected latent social factors and spatial adjacency. The spatial distribution of social areas with different latent dimensions is referred as a *social landscape* (Liu, 2014; Liu & Cao, 2017).

2.3. Deriving Activity-based Communities to Map Social Areas

One critical issue of social area analysis is that it heavily relies on social demographic surveys, i.e. census, which is not available for all countries. Census also often has a long interval, i.e., ten years, which cannot characterize rapid change. Socially sensed data, however, have the potential to fill the gap as the association between SES and activities derived from socially sensed data sets (e.g., Twitter, Foursquare) has been explored. However, such associations are still an open question especially as there are only certain case studies, and not all of them focus on modeling associations at city scale. Therefore, the first research objective of this study is to

investigate the influence of SES on human mobility at individual and collective level across a city by modeling the relationship between SES and human mobility metrics that are derived from empirical data for Washington, D.C.

This study further proposes an analysis framework that utilizes socially sensed activity data to derive neighborhoods, represented as groups of geographical units with homogeneous SES, which can be an alternative approach when detailed social demographical data are not available for social area analysis. Studies such as Cranshaw, et al. (2012) have tried to segment urban space into individual neighborhoods based on human activity signatures in the space denoted by the neighborhoods. There are also studies modeling human movements between neighborhoods as a complex network and employing community detection in network analysis to find out subnetwork structures, such as the inter-country mobility community using georeferenced tweets as a proxy for movement (Hawelka et al., 2014), or inter-neighborhood community using CDRs (Gao, Liu, Wang, & Ma, 2013). De Montis, Caschili, & Chessa (2013) and Šćepanović, et al., (2015) employed community detection on worker commuting networks from survey to delimitate municipalities into multilevel cohesive regions in terms of commuting activity.

This study focuses on grouping existing geographical units based on their interaction intensity that is defined by all human transitions between them, and further models the connection between the purely networked based communities to the classic SES based social areas. The underlying hypothesis is that *neighborhoods with strong activity interactions also share similar social areas*. Interactions are important because

the lack of cross-group interaction can induce prejudice as social psychologists suggest (Pettigrew, 2008). Previous studies as introduced in Section 2.1 suggest that people with similar SES live closer to each other. If residents in the same neighborhood share similar activity spaces and if they visit neighborhoods whose residents also have similar SES more frequently, their home neighborhood may have stronger connections to those neighborhoods with similar SES. If such a neighborhood group, referred to as an *activity-based community* in this study, can be derived from an activity interaction-based model and their spatial coincidence with neighborhood groups derived from social areas analysis can be confirmed, then these activity-based neighborhood groups can be used as an alternative to describe the social landscape of a city.

For this research, we select georeferenced tweets as a proxy of human activities as Twitter data is one of the few open accessible activity-related data sets. Even though georeferenced tweets are criticized for a bias towards younger, high-income and urban users (Malik, et al., 2015), these data are commonly used in human activity-related studies. Compared to CDRs, the volume of georeferenced tweets and its coverage of the population is smaller. However, the referenced tweets have more spatial detail so that they can be directly aggregated to existing geographical units delineated for census, i.e., census tracts or block groups, rather than using Voronoi tessellations derived from cellular towers' service areas as the proxy of a record's location.

Two major research objectives are addressed in this study. First, activity indicators and SES are derived from georeferenced tweets and from social area analysis, respectively. Their association is modeled and discovered. Second, census tract groups

with cohesive activity interactions are derived from a network representing the activity interactions between these tracts. The spatial coincidence of the groups derived from both activity and from SES are analyzed and their spatial matching is determined.

The rest of this chapter is organized as following. Section 2 introduces the study area, the data sets for modeling activity, and some key preprocessing steps to filter out required data records. Section 3 introduces the spatial distribution of human activities in the city. Section 4 describes the approaches and results to retrieve the spatial distribution of SES, the human mobility patterns and spatial interaction patterns, and to analyze their association. Section 5 concludes the main findings of the study and proposes future work that extends this research especially with respect to map social landscape in areas lacking of survey data.

3. Study Area, Data and Data Preprocessing

Washington, D.C. is selected as the major study area for this research. Washington, D.C. is the capital of the United States. This city has a very large number of commuters who work in the city but live in metropolitan areas in adjacent counties in Maryland and Virginia. Therefore, peripheral areas in the Washington, D.C. Metropolitan Statistical Area are also included as part of this analysis. The Washington, D.C. metropolitan area has been experiencing rapid growth over the past three decades (Sexton et al., 2013; Song, et al., 2016). Within the District of Columbia, neighborhoods are also undergoing continuous change in the form of gentrification that often leads to the displacement of residents (Jackson, 2014; Blessett, 2015). Census tracts are used as the basic geographical unit of analysis. Sociodemographic attributes are collected from

the American Community Survey 2011-2015 using the 5-year estimates. The D.C. area has a high population density although the majority of the residents live in its suburbs (Figure 2-1). One significant demographic characteristic of the Washington, D.C. metropolitan area is that White populations tend to live on the west side of the city and in the metropolitan area (Figure 2-2). More details about the socio-economic attributes will be discussed in later sections.

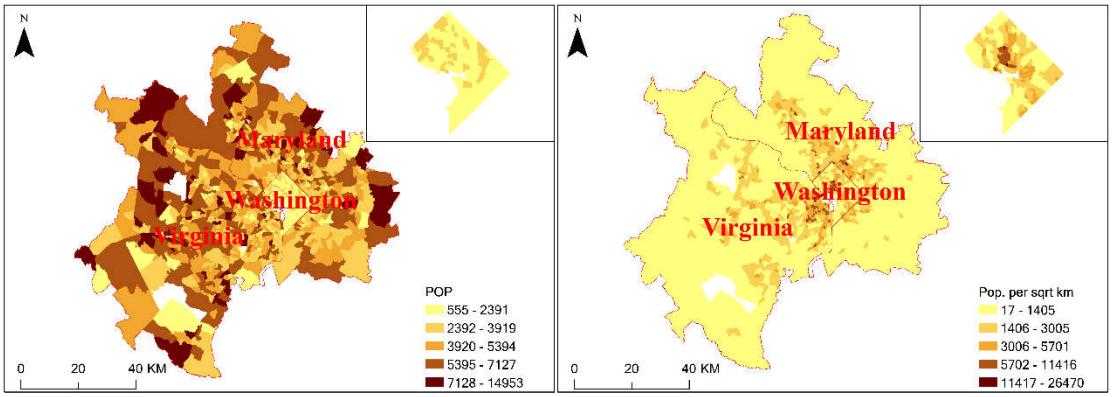


Figure 2-1 Residential population and the population density of Washington, D.C. metropolitan area in 2015 based on 2011-2015 5-year ACS.

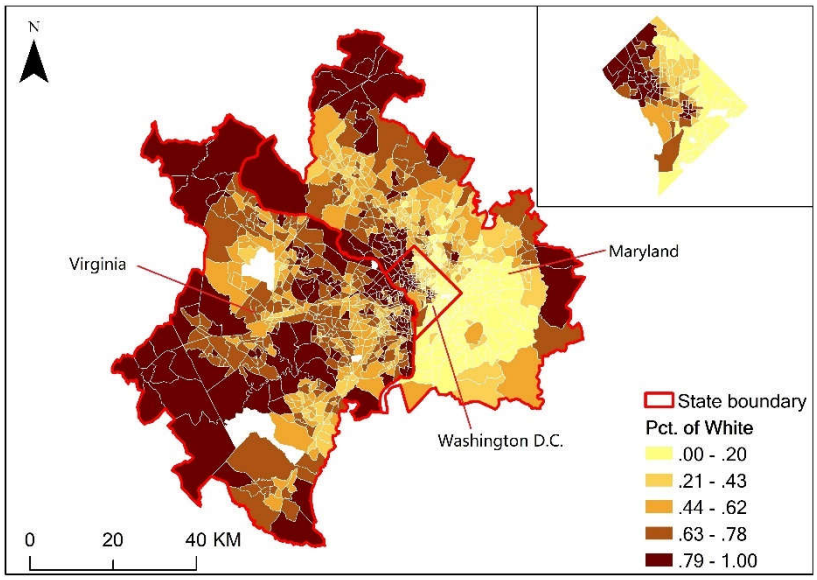


Figure 2-2 Percentage of White population in the study area.

Georeferenced tweets from the study area were collected for 380 days between April 2014 - 2016. Accounts that use location spoofing are identified and tweets from

these accounts are removed. Location spoofing is a technology that allows smartphone users and bots to use a false location instead of their real location while using location-based social media services. Although there is recent new technology that can identify many spoofing scenarios using a sophisticated strategy (B. Zhao & Sui, 2017), here only the cases where users with tweets that are only from one or two same GPS coordinates is addressed and removed, since counting the tweets being processed by fixed locations is a risk when evaluating the distribution of tweets in space. while scatted spoofing locations are not expected to have any extreme influence. For the data set in this study, about 8% of our collected georeferenced tweets are identified as possible spoofing situations. Accounts whose daily average tweets exceeds 40 are also removed. Accounts that may not be from local residents are also removed. The n-day rule (Li, Goodchild, and Xu, 2013; Hecht and Stephens, 2014; Johnson, et al., 2016) is applied to remove tweets from non-local people whose tweet footprint appear less than eight staying days in data collection period. As a reference, the median staying days of Twitter users who are observed in the tracts where the National Mall and Dulles International Airport are located are four and five, respectively. After the preprocessing, there are 5,317,420 tweets from 45,446 users remaining, representing 87% of the original tweets and 22% of all observed users.

4. General Spatial Activity Patterns

The density of the georeferenced tweets is heavy-tailed in mathematical forms. It best fits a truncated power-law distribution ($\alpha = 1.6, \lambda = 4.3$) when using a 100-

meter grid tessellation, and best fits a lognormal distribution ($\mu = 2.7, \sigma = 1.4$) when using census tracts as the geographical unit. Even if there is slightly difference between the best mathematical fitting, which may be subject to modifiable areal unit problems (MAUP Openshaw, 1984), the heavy-tailed distribution is consistent with respect to the grid tessellation and census tract divisions. Such a highly skewed distribution is also observed in previous studies on spatial distribution of human activities in cities (Jiang et al., 2016). It suggests that most activities crowd into a few places in the metropolitan area. The spatial distribution of the density (Figure 2-3) shows that the downtown area of Washington, D.C., such as the National Mall, Capitol Hill, the White House, etc., attracts most tweets. There is no correlation observed between the residential population and the tweets at the tracts level (Pearson's $r = -.0.01$, $p > 0.1$; Spearman's $\rho=0.13$, $p < 0.01$). This further suggests that there is moderated spatial mismatching between the spatial distribution of tweets and population at the census tract level by employing metrics to measure spatial segregation. The Index of Dissimilarity and Gini Coefficient (Massey & Denton, 1988) between the two variables is 0.52 and the Index of Dissimilarity (Iceland, Weinberg, & Steinmetz, 2002) is 0.38. Both metrics range from 0 to 1, where 0 represents the case where two types are evenly distributed in space, and 1 represents full segregation.

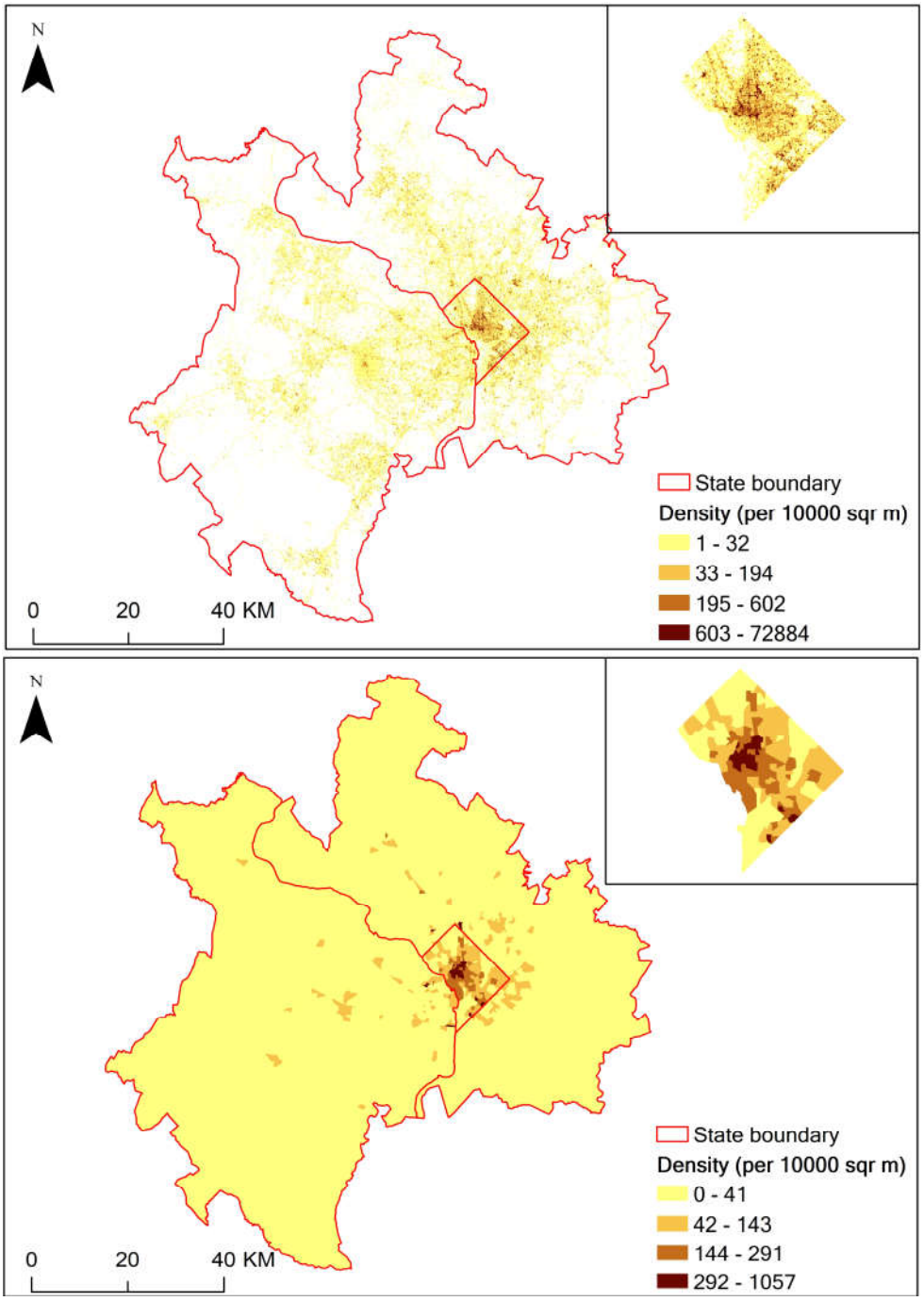


Figure 2-3 Density of tweets in the study area. a) by 100-meter grid b) by census tract. Color ramp breaks are based on a head/tail breaks classification (Jiang, 2013).

5. *Analyzing SES-activity Relationships*

5.1. Main Workflow

The workflow of this study consists of two main tasks (Figure 2-4). The social area, i.e. the landscape of SES distribution, is determined by applying PCA analysis on

sociodemographic variables using census data as ground truth. Human activities (i.e., movement patterns) are modeled at two scales. One scale captures aggregated individual mobility pattern over tracts, and the other capture people's transitions between tracts. For both scales, associations with SES are modeled, and the statistics of individual mobility by different SES groups are investigated. The mobility pattern for individuals from the tract groups that are identified by analyzing the structure of the tract interaction network, and are also associated with SES.

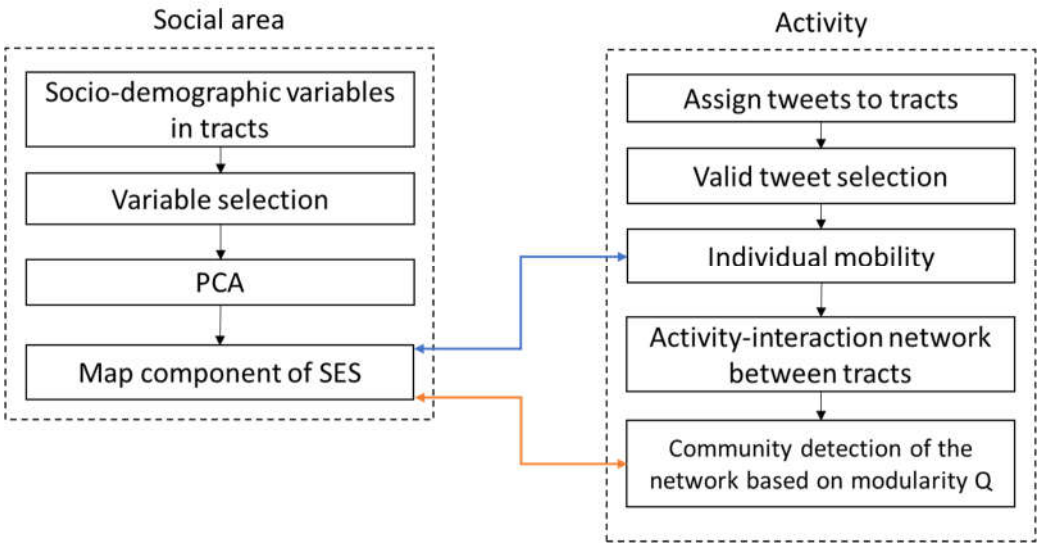


Figure 2-4 Workflow of the study

5.2. Derive SES of Census Tracts by Social Area Analysis

The candidate socio-demographic variables are from the same 79 variables used by Spielman & Thill (2008) with an additional 6 variables that describe population count by occupation categories. All candidate variables are numerical. A pre-processing step is applied to standardize some variables to percentage or density, for example, person-count related variables, such as population by gender, age, education, etc., are normalized by the population of the tract; household-count related variables such as

number of married households are normalized by the total households; household-unit related variables such as number of occupied house unit are normalized by the total household unit; monetary variables such as median household income and per capita income are converted to ranks. It should be noted that tracts with less than 100 people are excluded. As there are high collinearity between the candidate variables, a variable filtering method is applied to remove the variables with high correlation (>0.8) with the remaining variables. The remaining variables should also have a Kaiser-Meyer-Olkin score larger than 0.5. In addition, the Bartlett test of sphericity that tests whether the variable variances are equal across groups should also be statistically significant. The final selection is displayed in Table 2-1.

Table 2-1 Abbreviation and meaning for the selected socio-demographic variable

Abbreviation	Meaning	Abbreviation	Meaning
PCT_USCA	Percent of population under school age (< 5 years)	PCT_MWC	Percent of families married with children
PCT_SCHA	Percent of school age (5-17 years)	PCT_POVERT	Percent below poverty level
PCT_ELDER	Percent elders (>60 years)	PCT_UNEMP	Percent of workforce unemployed
PCT_FEM	Percent female population	PCT_CAR	Percent of occupied housing units with at least one vehicle
PCT_VACT	Percent vacant house	PCT_PUB	Percent enrolled in public school
PCT_OWOC	Percent owner occupied housing units	PCT_MINOR	Percent minority
PCT_HHCH	Percent households with children	MED_HHI_RK	Rank of median household income
PCT_ALONE	Percent living alone	MEDVALUEOO_RK	Rank of median value for owner occupied housing units
FAM_SIZE	Average family size	PERCAPIRA_RK	Rank of per capita income
PCT_MARR	Percent of families married	PCT_INCOME	Percent of gross rent in household income

For social area mapping, PCA with varimax rotation is selected as the mathematical tool to discover any latent factors that underlie the set of socio-demographic variables. Four factors with eigenvalues larger than 1 are addressed that

can explain 87.14% of overall variance. From the sign and weights of the top variables (Table 2-2), we are able to interpret themes for four latent components. Component 1 represents the dimension for *socioeconomic status*. Component 2 is more on *social rank*. Component 3 represents the *mix of social rank and SES.* Component 4 represents *economic* status.

Table 2-2 Selected principle components and top loading variables. Variables with absolute weight larger than 0.50 are displayed. The sign of weights indicates direction.

Component ID	Suggested theme	Percent variance explained	Top loading variables with weights
1	SES	39.05	PCT_MARR (-0.89) PCT_MINOR (0.89) MED_HHI_RK (-0.87) MEDVALUEOO_RK (-0.80) PCT_POVERT (0.67) PCT_UNEMP (0.60) PCT_MWC (0.58)
2	Social rank	25.25	FAM_SIZE (0.86) PCT_ALONE (-0.79) PCT_SCHA (0.70) PCT_PUB (0.64) PCT_CAR (0.62)
3	Mix of social rank and SES	16.51	PCT_HHCH (0.87) PERCAPITA_RK (-0.82) PCT_USCA (0.80)
4	Economic	11.32	PCT_OWOC (0.87)

For each tract, the original variable vector then is transformed by the four selected components so that the spatial distribution of the components can be further explored (Figure 2-5). In general, the components closely related to SES (Component 1 and 3) have a clear spatial clustering patterns. For people with similar SES, barriers based on physical geographic features (e.g., waterbodies or terrain) are not a major issue. It can be observed, for example, that the both sides of the Potomac River (the boundary between Virginia and Maryland) have similar SES. In the city of Washington, D.C., there is a clear spatial separation between residents with high and low SES, where the high SES residents cluster in the northwest while the low SES residents live in the east.

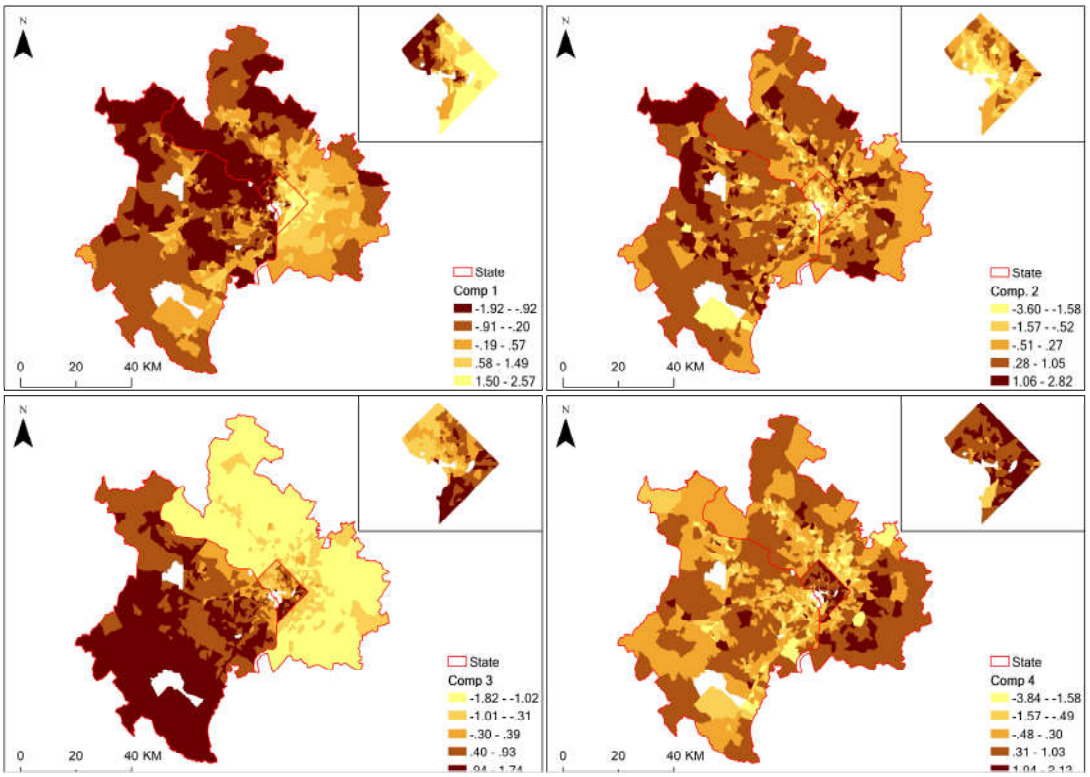


Figure 2-5 The geography of derived latent components. For the Component 1, a higher value indicates lower SES due to the signs of weights displayed in Table 2-2.

Since Component 1 explains nearly 40% of the total variance and its social area has a clustering pattern, this component is used as the proxy of SES landscape for further comparison with the community map derived from the activity-based network analysis.

5.3. Model the Relationship between SES and Human Activity

The mobility associated with undertaking daily activities can be characterized by different perspectives. *Radius of gyration* and *entropy* of visited tracts are employed in this study. Radius of gyration is commonly used to measure the spatial dispersion of an individual's daily activity (González et al., 2008; Song, et al., 2010; Hawelka et al., 2014; Zhao et al., 2016). It is formalized as:

$$r_g = \sqrt{\frac{1}{n}\sum_{i=1}^{n} |\bar{a}_i - \bar{a}_{cm}|^2}$$

where $\bar{a}_i$ is a location in an individual's daily travel and $\bar{a}_{cm}$ is the center of mass of the daily travel.

This is however, a subjective concept of "daily travel". As human behavior is known to be bursty (Barabasi, 2005), where people may post a few tweets in a short window and wait a long interval between two posting windows, it cannot be expected that the everyday tweets are evenly distributed along an individual's trajectory. Unlike the CDRs that continually track an individual's movement over space, georeferenced tweets are just a set of discrete snapshots, irregularly distributed along the time dimension. We assume, therefore, that people are more likely to post tweets in the places they regularly visit. To reconstruct regular daily travel based on places that are visited, DBSCAN (Ester, et al., 1996) is applied to filter out the spatial clusters of tweets as such places. For each place, the median time of day of tweets in a cluster is used as the time that the place is visited. All places are then sorted by these median times to form daily trajectories.

After that step in the workflow, each individual's radius of gyration is calculated based on two estimations of a "daily travel". One is to take account of each place only once, and the other is to weight a place by the number of tweets in that place, which is a proxy of the frequency of visits. Empirically, the Pearson's r correlation between the two results is 0.93. Therefore, only the unweighted approach is used in this study.

To measure diversity, entropy is widely used. In the context of measuring the diversity of places visited, this is formulated as:

$$H_u = -\sum_{i=1}^{T} p_{u,i} \log(p_{u,i})$$

where $p_{u,i}$ is the proportion of an individual u's tweets that are observed in a census tract. Larger values of entropy indicate a diverse activity pattern.

It is also necessary to infer an individual's home tract so that the derived SES can be linked to the individuals' observed activity patterns. Similar to Xu et al. (2016), we select tweets that are posted at night (between 21:00 and 7:00 of the next day) as the candidate tweets posted from home. The census tract that has the most unique dates among the candidate tweets is selected as the home tract of the individual. After this process, the mobility metrics of an individual can be linked to the census tracts. We were able to identify home tracts for 41,645 individuals', accounting for 91.6% of all 'local' users.

The overall distribution of radius of gyration and entropy for visited tracts are highly skewed with the distribution of radius of gyration fitting an exponential distribution, while the entropy fits a power law distribution best. The thin tail of the radius of gyration distribution, which is slightly different from the observed long-tail distribution reported in previous studies, may be due to the geographical boundary of the study area that excludes long-distance trips. In addition, similar to the overall tweet distribution, the identified homes among tracts are also highly skewed with the median identified home counts in tracts being 15, and approximately 25% of the tracts have

less than 11 Twitter users whose radius of gyration can be identified. This unevenly distributed sample leads to data sparsity in these places.

The relationship between SES and the mobility metrics are complex. People with high SES do not necessarily have a larger spatial range than the people with low SES (Figure 2-6a). Figure 2-6a also suggests that the diversity among the people with the same high SES is very large as the standard deviations are larger than the means. This may due to the mismatch between the census-based SES indicators and the diversity of people living in the same tract or household. For instance, different family members may have different activity patterns from each other, but all of them are categorized into the same SES group. In addition, groups with higher median household income living in the city may have a smaller range of mobility than populations with similar income levels who are living in suburban tracts, as they have very different lifestyle and transportation modes, e.g., public transportation and bicycles in the city *versus* vehicles in suburban areas. By using the median value of individuals as a *typical* individual in the same tracts, it can be observed that the average radiuses of gyration and their standard deviations decline from high SES tracts to low SES tracts (Figure 2-6a). For the entropy of visited tracts, people with lower SES have slightly higher diversity as the values of entropy incline with SES class numbers (Figure 2-6b). This may be due to the spatial segmentation of census tracts where the tracts in the city are much smaller than those in suburban tracts, and the density and accessibility to public transportation inside cities like Washington, D.C. may be better than that in suburbs. The influence of residential location on the SES-mobility relationship is also supported

by the by-state mobility metrics in Figure 2-7 where individuals in D.C. have smaller spatial activity dispersion but higher activity diversity.

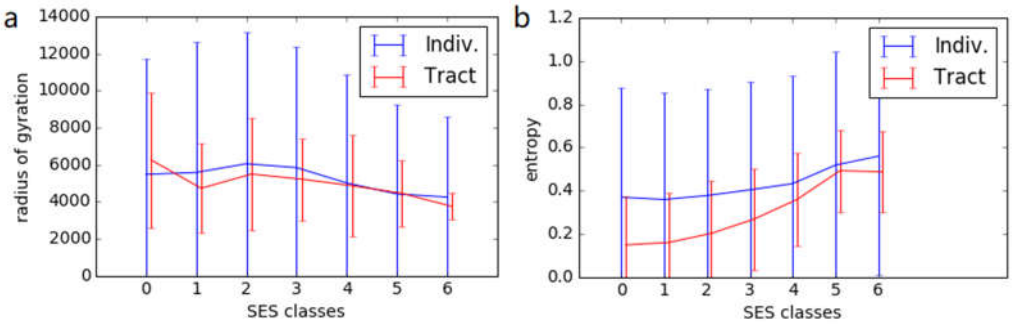


Figure 2-6 a. Blue: the radius of gyration in individual groups with different SES. Red: the radius of gyrations in tract groups that individuals' radiuses of gyration are aggregated to tracts and the median of aggregated values are used as the representative value of the tract and the tracts are categorize by SES. b. Blue: the entropy of visited tracts in in individual groups with different SES. Red: the entropy of visited tracts in individual groups with different SES. For all cases, SES groups are categorized by equal intervals on values of Component 1. Larger class number indicates lower SES.

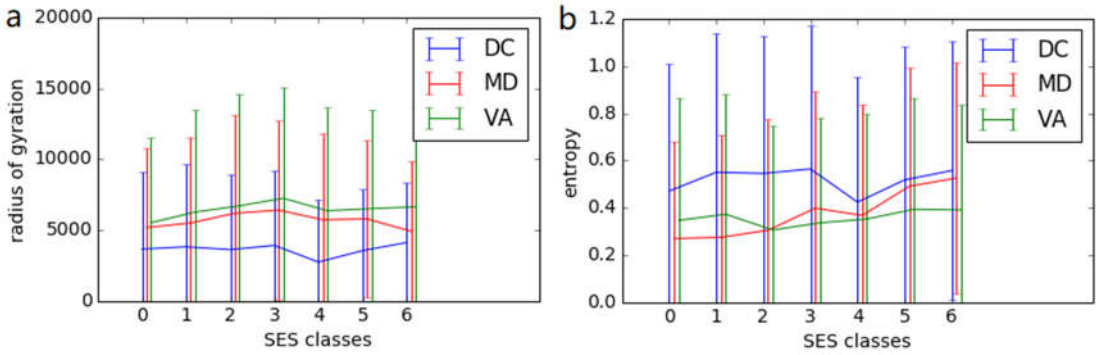


Figure 2-7 a. The radius of gyration in individual groups with different SES by states. b. The entropy of visited tracts in individual groups with different SES by states.

5.4. An Activity-interaction based Data-drive Approach to Infer Regions with Homogeneous SES

From empirical analysis, it can be concluded that inferring a single tract's SES from its residents' mobility pattern is difficult, as the relationship between SES and basic mobility metrics is complex. In addition, some tracts have sparse data with respect to activity metrics. An alternative approach is to infer groups of tracts that have similar

SES and spatial connected as neighborhoods by modeling the interactions between the tracts. For these tracts with data sparsity, they may be commercial areas or peripheral tracts with few residents. However, these places still may be sources and targets of movements. Therefore, modeling the interaction between the tracts can assist with assigning them to neighborhoods that match their social functions.

It can be observed that census tracts with similar SES are also spatially contiguous and aggregated, for example, the derived Component 1 in Washington, D.C. (Figure 2-5). Such socially cohesive and spatially connected groupings can be derived by spatial clustering approaches, such as LISA (Anselin, 1995) from a geostatistical perspective and k-means and spectrum clustering from a data mining perspective (Han, Kamber, & Pei, 2012). LISA is used to return the statistically significant spatial groups (Figure 2-8).

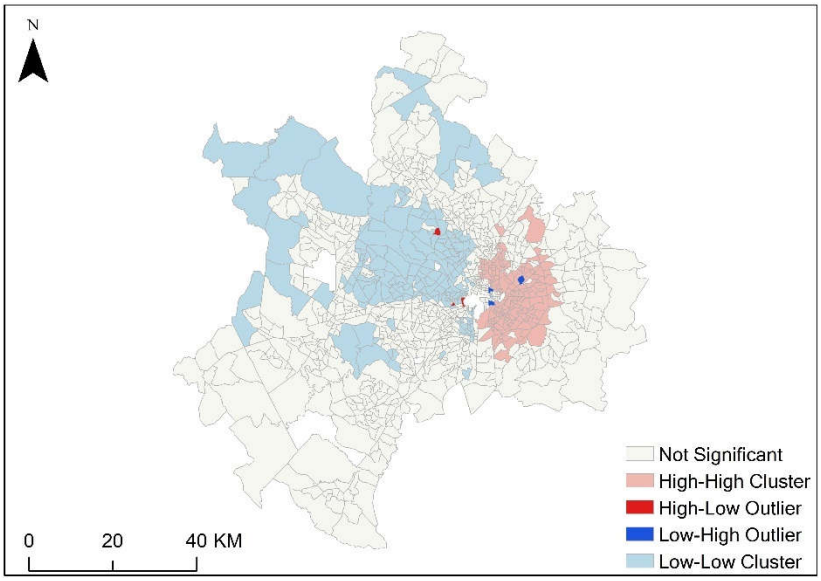


Figure 2-8 Spatial clusters of Component 1. Due to the sign of the component weights, the High-High clusters are the neighborhoods with low SES compared to their neighbor tracts and the Low-Low neighborhoods clusters represent neighborhoods with high SES.

Residents in the same tract may share similar daily activity spaces, even if individuals' mobility varies significantly. Such sharing can be represented as human flow between the tracts. Therefore, if activity space groups can be derived from the pattern of tract interactions similar to the groupings derived from social area clustering, such groupings may still be helpful for understanding the SES landscape with spatial detail for areas that are without detailed sociodemographic surveys.

The interactions between census tracts is modeled as an undirected network and represented as a network graph G = (V, E), where G is the graph; V is the set of census tracts as nodes in the graph; and E is the set of links between a pair of nodes in V if there are interactions between them. In this study, *interaction* is defined as the cooccurrence of a Twitter user in both tracts. The strength of the interaction between tract *i* and *j* is thus defined as the number of users who appears once in both of the tracts, as also used by (Lansley & Longley, 2016). It is denoted as I_{ij}, where $i \neq j$, meaning there is no self-loop in the network graph. Unlike De Montis, Caschili, & Chessa (2013) and Šćepanović, et al., (2015), the transition between any tract pair is not necessarily part of a daily commuting trip as third-place visiting is also an important part of the daily activity that characterizes lifestyles. In addition, it is also not necessary to assume that it is a user's movement from one tract to another is a complete trip or a pass-by. However, such a transit suggests that the user is aware of the physical and social landscape of both places.

5.4.1. Structural Characteristics of the Activity-based Tract Network

It is important to test the hypothesis that if the strengths between any pair of two tracts are evenly distributed in the network as a homogeneous network has no group structure inside. The frequency distribution of the interaction strength I_{ij} shows that the distribution is highly heterogeneous as shown in Figure 2-9. A strong influence of locality is also explored: strong interactions are most likely from adjacent tracts and the strongest inter-tract interactions all happen between adjacent tracts. This can be the result of two factors, i.e., from the behavior modeling perspective, it is easier to visit nearby places, which makes the inter-tract interactions strong, and from the data quality perspective, these interactions may also be influenced by the uncertainty of GPS positioning of smartphones (Zandbergen, 2009), especially at the border of tracts.

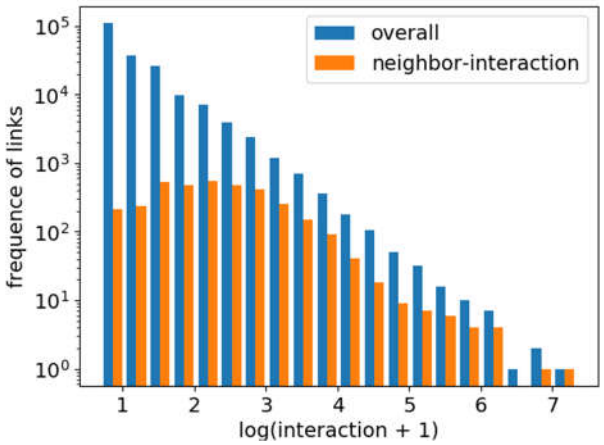


Figure 2-9 The frequency of interaction strengths between any pair of tracts in the graph, and the interactions between adjacent tracts.

Node centrality is another important property for a network. *Betweenness centrality* is commonly used to identify the hub tracts in the activity network (Barthélemy, 2011; Gao, et al., 2013). *PageRank* score (Page, et al., 1998) was initially used to measure importance of webpages but it is also a general form to identify the

accessibility of the nodes taking accounts all links and weights (Zhong, et al., 2014). Both metrics have a similar spatial pattern (Figure 2-10 and Figure 2-11) that the public space, such as major commercial places (e.g. Tysons Corner, an edge city and shopping center in DC area), traffic hubs (e.g. the Dulles Interactional Airport), and recreation places (e.g. the National Mall and its surrounding places) have a high betweenness centrality and PageRank scores. Some exceptions include highly compact cities, such as Silver Spring, MD, where the major residential land uses are multi-level apartments mixed with commercial uses. This suggests that these public spaces, which are workplaces (and are third places), or the overlay of both, rather than the residential areas, are the major places where interactions happen in the Washington, D.C. area.

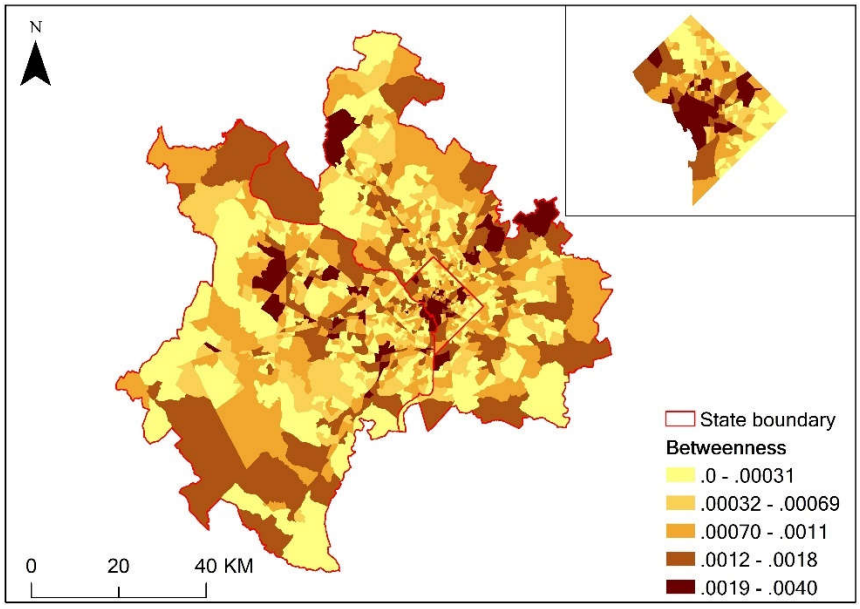


Figure 2-10 Tract betweenness in the network.

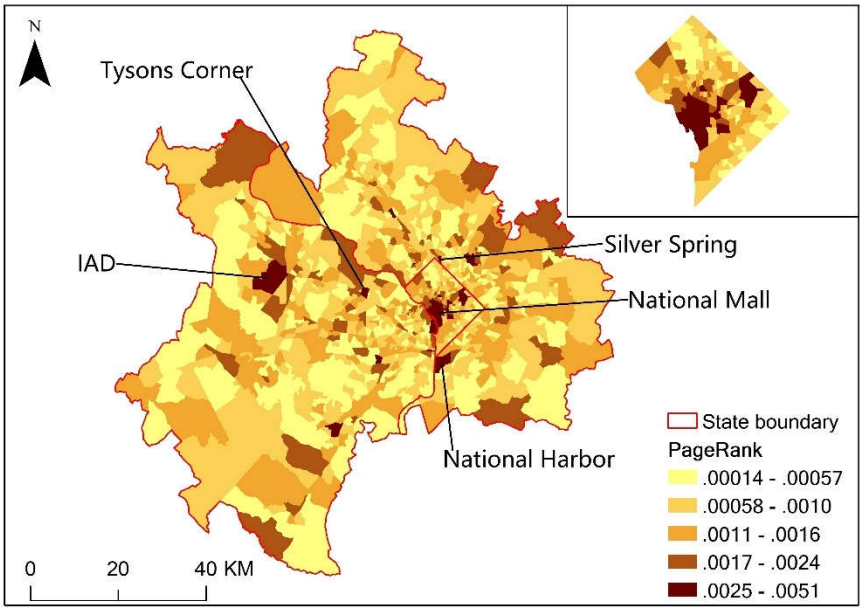


Figure 2-11 PageRank score of the tracts. IAD is the Dulles International Airport.

5.4.2. Community Detection

Community detection relies on two main approaches, hierarchical and partitioning-based (Girvan & Newman, 2002). In this study, we employ a well-known partitioning-based method (Blondel, et al., 2008) by optimizing the *modularity* (Newman, 2006) of the subgraphs. Modularity compares inner community links of a partitioning solution to a null model where all links are randomly assigned. The modularity of a weighted network is formalized as:

$$Q = \frac{1}{2m}\sum_{i,j}(A_{ij} - \frac{k_i k_j}{2m})\delta(c_i, c_j)$$

where Q is modularity; m is the size of edge set E in an undirected network; A_{ij} is the weight of the link between node i and node j; k_i is the sum of weights of node i, and C_i is the community to which node i is assigned; and the value of δ function is 1 if $c_i = c_j$ and 0 otherwise. By optimizing the modularity, a network partitioning resolution makes the intra-community connections dense and inter-community connections sparse.

To reduce the strong locality as mentioned in the previous section, all links between adjacent tracts are removed from the original network. In addition, as the overall interaction is highly skewed (Figure 2-9), the logarithm of the original link weight is used as the weight. Following a knowledge discovery approach and similar to previous studies (e.g., De Montis et al., 2013), a hierarchical partitioning community detection process is employed where each community detected from the first round partitioning (denoted as Level 1 communities) is further partitioned as an individual network by the same partitioning algorithm (denoted as Level 2 communities).

The spatial pattern of the first-round community detection results shows the evidence of integrated influences of locality, physical geography, and socio-demographic factors (Figure 2-12). Even if the links between adjacent tracts are removed, almost all community members are spatially contiguous and adjacent. Almost all tracts in Virginia are assigned to the same community (Community 1). The boundary between Community 1 and 2 is the Potomac River, the boundary between Maryland and Virginia. The limited transportation corridors on the river could be the reason for such a clear separation, even if suburban residents are highly mobile with vehicles. This can also explain the boundary between Community 1 and Community 4. Within D.C., the Anacostia River runs inside the city of DC separating the southeast region from the city, while there is no physical barrier between D.C. and Maryland, so that some D.C. tracts have stronger interactions with tracts in Maryland and these are assigned to a Maryland-based community (Community 4). Tracts in Maryland are mainly assigned

to either Community 2 or 4. The spatial extent of Community 4 has a similar shape with a cluster reflecting a high percentage minority population (Figure 2-2).

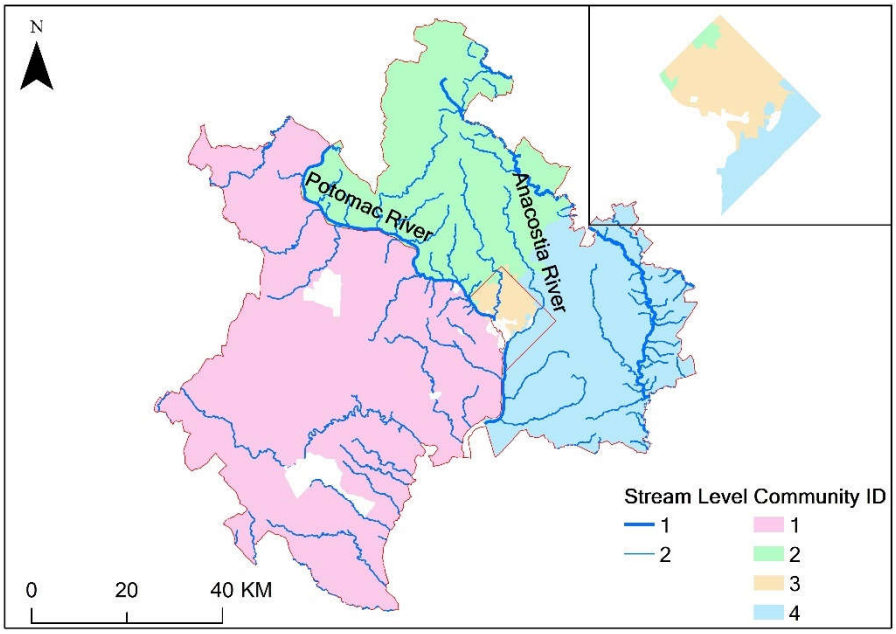


Figure 2-12 Spatial pattern of communities detected by the first-round community detection (Level 1 communities).

The spatial distribution of the Level 2 communities are not neccesaritly bounded by spatial contigency (Figure 2-13). There are 23 Level 2 communities in total that each Level 1 community does not necessarily have the same number of child communities. Some communities have enclaves spatially surrounded by other communties, such as Community 32 in DC, and Community 43 (The first digit of a Level 2 community is correponding to the same ID of their parent community in Level 1). Therefore, SES appears to serve as a major influence at this level as the enclaves might be the result of activty preferences based on selecting third places influenced by the SES of people moving among these tracts.

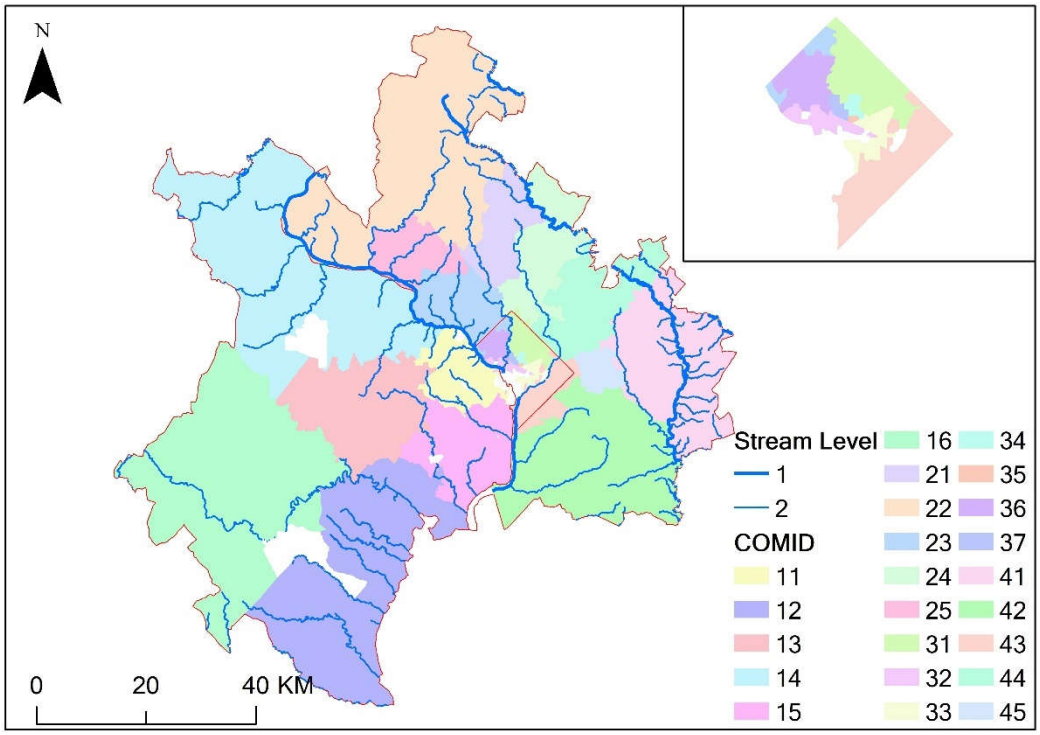


Figure 2-13 Spatial pattern of the communities detected by the second-round community detection (Level 2 communities). The first digit of the community ID is corresponding to the community ID in Figure 2-12.

5.5. Neighborhoods Comparison

The measurement of the agreement between the neighborhoods derived from SES clusters and activity-network communities consists of two perspectives: 1) spatial segmentation; and 2) relative SES ranks. An ideal match would mean that both approaches aggregate the tracts with the same pattern, and that differences in activity patterns derived from the activity communities can also differentiate their SES.

V-measure (Rosenberg & Hirschberg, 2007) is employed to measure the agreement on the two spatial segmentation solutions. V-measure is an entropy-based metric ranging from 0 to 1 where 1 represents a perfect match. The value of V-measure between the clusters from SES LISA result and the tract communities derived from activity-based network is 0.27. In addition, the two solution results are separately

evaluated using individual values that represent each state. The values for Washington, D.C., Maryland, and Virginia are 0.34, 0.33 and 0.17, respectively. This suggests that the overall agreement of the two segmentation solutions are moderated, and disagreement is influenced by geography. It should also be noted that the LISA clusters are statistically significant while the communities detected from the activity-based network do not follow such a restriction. In addition, if the study region was the tracts in Washington, D.C. only and both algorithms are applied to this subset of tracts, the V-measure of the new resolutions raises to 0.48. Therefore, the SES clusters and activity-based community has good agreement in Washington, D.C. This may due to the fact that many activities happen in the core of the metropolitan area and that the association between activity patterns and SES are more significant there than in the periphery of the city.

We undertook one more analysis using the measures that we calculated. The findings presented in Section 5.3 show that lower SES tracts have smaller radii of gyration and smaller standard deviations, however they exhibit larger entropy of visited tracts. Given these results, we decided to test how a compound indicator performs as a proxy of SES: $\frac{\overline{H_c}*\overline{Med(r_g)_c}}{Std(Med(r_g)_c)}$, where $\overline{H_C}$ is the mean of individuals' entropy of visited tracts in the tracts that belong to a community *c*. $\overline{Med(r_g)_c}$ is the mean of the medians of individuals' radii of gyration in the tracts within a community *c*. $Std(Med(r_g)_c)$ is the standard deviation of the tract medians. For the clusters and communities, this compound indicator returns a 0.47 on Pearson's r, and 0.55 on Spearman's ρ using the

mean of SES values. These results are better than using the activity metrics as a proxy for SES.

6. *Conclusions and Future Work*

This study employs georeferenced tweets as a proxy of residents' activity (especially travel movements) in a US metropolitan area. It is observed that the majority of activities are focused in the downtown of the core area (i.e., downtown Washington, D.C.). Further exploration of the association between SES and activity shows that the mobility of individuals is influenced by the SES but it is not a simple universal correlation as observed by previous studies. Tracts with low SES residents have lower spatial dispersion but have higher activity diversity in space. This may be due to the urban geography of the Washington, D.C. area where individuals with low SES live in the downtown area and the city has highly accessible public transportation. In addition, it is also observed that people living in the suburbs have higher spatial dispersion but lower activity diversity than those who live in downtown with the same SES. This may also be due to urban geography that the suburban areas have lower density of commercial and residential places. By analyzing the interaction network between the tracts, it is observed that physical geography, especially rivers, continue to play an important role in shaping people's movement over space even if residents have high mobility with vehicles.

Concerning the observed complex relationship between SES and residents' mobility patterns, we applied a data-driven approach that models the interaction between the tracts and further associates the tract groups derived from the interaction

patterns to infer SES, rather than inferring tracts' SES by their residents' human mobility directly. By partitioning activity-based interaction network and using the residents' mobility as the proxy of SES, the methods are still able to derive the neighborhoods with different SES to capture the social landscape of the study area. The matches on spatial segregation and socioeconomic characteristics between the two approaches have moderated agreement.

There are several limitations that could be improved in future studies. Due to the spatial bias of Twitter users, collecting additional socially sensed data in the periphery would be useful. In addition, the current activity-based tract-interaction network does not calibrate the influence of physical geography. The individuals are modeled as being identical in that their SES are assigned by the SES of the whole tract. However, SES for residents should be variable even in the same tract due to varying ages, genders, lifestyles, etc. Such information might be inferred by the text content of their tweets so that the model can give more detail on the influence of SES to different groups.

Chapter 3: Identifying Spatiotemporal Urban Activities through Linguistic Signatures

1. *Abstract*

Identifying the activities that individuals conduct in a city is key to understanding urban dynamics. It is difficult, however, to identify different human activities on a large scale without incurring significant costs. This study focuses on modeling the spatiotemporal patterns of different activity types within cities by employing user-contributed, geosocial content as a proxy for human activities. In this work, we use linguistic topic modeling to analyze georeferenced twitter data in order to differentiate different activity types. We then examine the spatial and temporal patterns of the derived activity types in three U.S. cities: Baltimore, MD., Washington, D.C., and New York City, NY. The linguistic patterns reflect the spatiotemporal context of the places where the social media content is posted. We further construct a method to link what people post online to the activities conducted within a city. We then use these derived activities to profile the characteristics of neighborhoods in the three cities, and apply the activity signatures to discover similar neighborhoods both within and between the cities. This approach represents a novel activity-based method for assessing similarity between neighborhoods.

2. *Introduction*

Urban life involves a variety of activity types that are an intrinsic part of urban dynamics, including commuting, shopping, dining out, etc. Exploration and analysis of

these different types of activities leads to a better understanding of the pulse of the urban landscape, e.g., transportation, economic, and social behaviors. People's activities in the street comprises Jane Jacobs' "sidewalk ballet" (Jacobs, 1961). Activities also help to delimitate *places.* From *structuration* theory, places are established only if they are locations of constant and reiterative activity (Cresswell, 2014). Poststructuralist *assemblage* theory that refers to the emerge of new unique wholes from the interactions between parts also highlights that the dynamics in a city contribute to an emerging sense of place (Dovey, 2012). Therefore, understanding differences in activity types, and the magnitude of these activities at different locations in a city provides information on the intrinsic nature of different places. Sensed activities can be utilized for decision-making in urban planning or for improving services.

One conventional method for characterizing parts of a city, i.e., neighborhoods, is to use demographic data. For example, the ESRI Tapestry[2] project categorizes residential neighborhoods in the United States into 67 types by employing Census data. Census data, however, does not reflect how people actually interact with urban spaces, and does not cover the socio-economic aspects of the neighborhoods that incorporate, for example, commercial areas, since a census only surveys residents. Using a derived activity distribution among the neighborhoods, we can categorize neighborhoods from an activity-based perspective, and compare the similarity of neighborhoods based on this new perspective.

2 http://www.esri.com/landing-pages/tapestry

Sensing human activities in a city can be financially expensive and time consuming. Given the complexity of modern survey techniques, researchers in different fields often survey only a sampled group of individuals with some denoted types of activity that are closely related to their study theme. For example, studies in transportation mainly utilize transportation activity surveys such as the U.S. National Household Transportation Survey (NHTS, Cervero & Kockelman, 1997; Chalasani, et al., 2005) or equip a limited number of enrolled vehicles with GPS loggers to track vehicle movements (Wolf, Guensler, & Bachman, 2001). Studies on public health also utilize travel surveys, for example, to link eating activities with a geographical context (Kestens, et al., 2010; Widener, et al., 2015).

Recently, socially sensed geospatial data sets (*social sensing*, Liu et al., 2015) have been used as proxies of human activities. Socially sensed geodata includes geographic information that are voluntarily contributed by individuals (volunteered geographic information, VGI, Goodchild, 2007), such as the geospatial data of OpenStreeetMap (OSM), georeferenced accident reports on Waze, and geospatial data that is collected but not purposely contributed by the individuals who generate the data (McKenzie & Janowicz, 2014), such as georeferenced taxi trajectories, call detailed records (CDRs), check-in (Cranshaw, et al., 2012), and georeferenced microblog posts from Twitter, a social network service (SNS). A georeferenced Tweet is a short message (typically text-based) limited to 140[3] characters from a Twitter user that includes metadata such as a location and a timestamp. In this work, we show how these tweets

[3] The character limit changed to 280 after September 2017

can be used to represent activities that are being undertaken by individuals in multiple cities.

Previous studies that utilized Tweets as proxies for human activities typically only model *posting a Tweet (tweeting),* as an identical activity, and used the variation of tweet volume only to characterize the social function of a region without fully utilizing the text in tweets that may provide further detailed activity type information. Projects, such as UrbanTick[4] by Neuhaus, relied on a change in the volume of tweets (spatially and temporally) to characterize the activity rhythm, or "the pulse of the city" (Michael Batty, 2010b). Such variations in tweet volumes are also used to characterize regions' social functions in a city by combining machine learning approaches (Wakamiya, Lee, & Sumiya, 2011; Frias-Martinez, et al., 2012; Lee, Wakamiya, & Sumiya, 2012).

The textual content of a tweet contains useful, descriptive information that is often overshadowed by the spatiotemporal meta data. Within the content of a tweet, people often explicitly or implicitly express their thoughts and feelings related to activities they are conducting when they are tweeting. Text analytics can thus extract place references and meaningful information from georeferenced tweets and construct place characterizations (MacEachren, 2017). One approach that has been taken previously is to filter related tweets by keywords, for example, Tsou et al., (2013)'s analysis on candidate names in the 2012 U.S. Presidential Election and Yang et al., (2016)'s system for exploring human dynamics based on people's interests.

4 http://urbantick.blogspot.com/2010/01/new-city-landscapes-interactive.html

Keyword analysis, however, may only expose specific events that involve a limited set of keywords closely related to the event. There may be new terms created to refer to a new event or a new type of activity that cannot be identified by a predefined set of keywords. Alternatively, an approach such as topic modeling that derives latent topics in text by a word-based statistical modeling approach can be used for knowledge discovery without predetermined keywords (Hofmann, 1999).

One of the most prevalent topic-modeling approaches is Latent Dirichlet Allocation (LDA, Blei et al., 2003). LDA assumes that each document in a corpus is associated with numerous latent topics that can be characterized by a unique word probability distribution. LDA and its variants on classification (Blei & McAuliffe, 2008; Ramage, et al., 2009) have been used extensively in previous spatial and place-based research (Adams, McKenzie, & Gahegan, 2015; Chae et al., 2012; B. Hu & Ester, 2013), but the standard LDA approach is arguably not a good model for tweets, given the limited text length in a typical tweet. One solution is to aggregate tweets as one long document based on locations or time intervals to fit into the standard LDA model (Eisenstein & O'Connor, 2010; Jenkins, et al., 2016; McKenzie, Adams, & Janowicz, 2015; Mehrotra, et al., 2013; Puniyani, et al., 2010). As alternatives, Twitter-LDA (Zhao et al., 2011) and Single Topic LDA (ST-LDA, Hong et al., 2016) assume that: 1) only one topic is involved in each tweet post due to Twitter's length limitation; and 2) multiple authors are involved in writing a collected tweet dataset. Such assumptions are similarly reasonable for this study, and for this reason ST-LDA is used as the primary means for topic modeling as it has also been applied to analyze resident-

government communication pattern in disaster (Hong, et al., 2017). Besides LDA models, Deep Learning frameworks on topic modeling have also been applied to the same task (Wang et al., 2016).

This research uses the volume profile of different activities as a quantitative means to retrieve knowledge about and the sense of places. This research explores the value of using a large user-contributed georeferenced dataset as a proxy for activities within and between cities on the east coast of the United States, and identifies and compares regions with respect to their activity profiles over several months. Using ST-LDA to build the model that links tweets to activities allows us to explore how activities are distributed both in time and space. This distribution can help us in two ways: First, the temporal and spatial patterns are used to validate the accuracy of the topic model in representing meaningful activities. Second, the overall distribution of the topics is employed to characterize places, such as different neighborhoods. The new computational model also provides feasibility to analyze the activity patterns with finer granularity in time and space as there is no pre-processing geographical or temporal units for aggregating the tweets to form a long text for fitting into a standard LDA model.

In this study, two major research objectives are addressed:

RO1. An natural language processing (NLP) workflow is applied to derive meaningful activity types from a large number of Twitter posts, and the resulting activity types are evaluated based on their spatial and temporal distributions. We specify a null hypothesize (H1) that the topics derived

from georeferenced tweet are identically distributed in space and time. In this work we will demonstrate how this null hypothesis is falsified.

RO2. The derived activities are used to profile the activity signatures of neighborhoods in three U.S. cities as a novel approach to characterizing the neighborhoods. The activity signatures are further employed to find similar neighborhoods both within and between cities. We specify a null hypothesis (H2) that aggregated topics, as proxies for activities, offer identical signatures that cannot differentiate one neighborhood from another. In this work we will nullify this hypothesis by showing that there are statistically significant differences in the topic signatures.

The remainder of this paper is organized as follows: Section 2 introduces the Twitter dataset collected from three cities in the U.S. for an empirical study. Section 4 discusses the approach used to extract activities from text in Tweets, and validates the set of derived topics via their spatio-temporal distributions. Section 5 shows how the neighborhoods are characterized by the derived activities and how the similar neighborhoods are found. Section 6 takes a neighborhood in Washington D.C. as a case study to show the effectiveness of the model presented in Section 5. The conclusions are presented in Section 7 of this paper, along with a discussion addressing potential limitations, and suggestions for future work.

3. Data

Twitter allows users to register anonymously and to post messages, labeled *tweets*, with rich metadata, including a unique ID for the tweet, a user ID identifying

which user posted a message, a time stamp indicating the time when the message was posted, to name a few. Within the content of a tweet, a user can use a hashtag (#) as the prefix to highlight a keyword to summarize the theme of the message or to draw others' attention. If users post tweets from a location-embedded mobile device, Twitter also allows users to include the device's coordinates as part of the tweet's metadata. Twitter provides a set of freely-accessible Public Streaming Application Program Interfaces (APIs)[5] that allows researchers to collect a sample of tweets in real time. Researchers can designate a specific region as a parameter to the API and collect georeferenced tweets from that area. Given a small enough region, it has been reported that almost all georeferenced tweets can be retrieved (Morstatter et al., 2013). This indicates that collecting data via the API provides a representative sample of the population of georeferenced tweets.

In this study, we collected georeferenced tweets from three U.S. East Coast cities: City of Baltimore (BC), Washington D.C. (DC), and the City of New York (NYC) as study areas. These three cities have their own unique socio-economic profiles: NYC is the largest city in the United States (population 8.5 million in 2015). DC is a smaller city (population est. 660 thousand in 2015) and the U.S. capital, known for its political activities. Baltimore, MD (population 623 thousand in 2015) is commonly identified as a city with a shrinking population.

Tweets were collected for these regions and filtered by a preprocessing step to remove the tweets from accounts that potentially use location spoofing. Location

5 https://dev.twitter.com/streaming/overview

spoofing is a technology that allows mobile device users to replace their real location by a predefined false location while using location-based SNS. Since location spoofing typically uses one false coordinate pair, a naïve rule is employed to remove tweets from users whose tweets are only from one or two same coordinates, although there is recent new but more complicated technology that can identify many spoofing scenarios with sophisticated strategies (Zhao & Sui, 2017). This cleaning filtered out approximately 8% of the data set. After filtering, 1,126,914 tweets remained for BC from October 2014 to April 2016; 1,737,225 tweets for DC over the same time period, and 5,234,725 tweets from NYC from February to August 2013. Although tweets from NYC are from a different time period to those collected for BC and DC, we do not believe this significantly affects the outcome of our analysis as the daily activity patterns in most parts of a city do not change dramatically over the span of a few months.

4. Methodology

The methodology section consists of two main steps. Section 3.1. develops a NLP workflow to derive topics from georeferenced tweets. The semantic meanings of the topics are investigated. Section 3.2. validates that the georeferenced topics can be used as proxy of activities and that their spatial and temporal profiles match the activities that are referred.

4.1. Extract Activity Topics from Georeferenced Tweets

In this section, we introduce the workflow to process tweet text using NLP tools and extract activity-related topics from processed text using ST-LDA topic modeling.

4.1.1. The ST-LDA Model

As is the case with many LDA approaches, ST-LDA treats words in a Tweet as discrete signals and utilizes the word frequency distribution among Tweets as statistical features (referred to as the *bag-of-words* model). ST-LDA assumes that each Tweet involves a latent *topic* underlying the words. Each topic is characterized by a unique probability distribution of the vocabulary that is used in the set of Tweets. That is, different topics have the same vocabulary but have different weights on words, which differentiate one topic from the others. One topic can be found in a group of Tweets with similar themes. The ST-LDA model can be treated as a dimension reduction method that maps Tweets from a very high dimensional vocabulary space to a relatively low dimensional topic space, while providing individuals with a way to interpret semantic meaning of each topic by exploring its word weights.

4.1.2. Natural Language Processing Pipeline on Data Preprocessing

A bag-of-words model presents a document as a vector whose indices refer to words, and values of items that refer to the frequency of the corresponding word in the document. Before simply splitting sentences into words and counting the frequency, there are several additional preprocessing steps required to clean the data. First, not all words in a tweet's text are informative. We prefer to keep words with meaning (noun, verb, adjective, etc.) that refer to entities, activities, movements, etc., rather than prepositions, determiners, and other words that likely do not refer to meaningful entities. Furthermore, the text contains phrases that should be treated as one entity rather

than separate words. In the typical bag-of-words model, each word in a phrase is treated as an independent unit. For example, “New York” may be processed as “New” and “York”, which does not reflect that the original phrase is referring to a certain entity. Phrases must be explicitly denoted in the bag-of-word model. Phrase detection was used to bind the words in a phrase through the use of an underscore, e.g., “New York” is represented as a single token “New_York”. In addition, standard stop words were removed for clarity and to save computational time in further NLP steps, which is also a standard step in most NLP models. The data cleaning step is outlined in *Figure 3-1*.

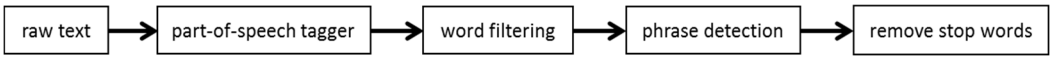


Figure 3-1 Word preprocessing

This preprocessing workflow first uses the Ark Twitter Tagger (Gimpel et al., 2011) to split sentences into independent words (referred to as tokens) and tag the part-of-speech for each token. The part-of-speech is a category, to which a word is assigned a label based on its syntactic function in a sentence, such as common noun, proper noun, verb, etc. Then, word filtering is applied to keep words referring to entities, such as nouns and verbs. The remaining tokens are processed by an NLP package Gensim’s (Řehůřek & Sojka, 2010) phrase detection module. In the final step, the stop-word removal is applied to each sentence based on an English stop word list in NLTK (Loper & Bird, 2006), which is a common natural language toolkit.

4.1.3. Interpreting Topics

The ST-LDA model assumes that a set of tweets involves *T* latent topics, and each tweet has at most one non-noise topic. The output of this approach includes *word*

distributions for each topic and *labeled tweets*, whose label is a unique ID referring to one of the extracted topics. The word distributions of sample topics are displayed as word-clouds in Figure 3-2. Each topic's theme can be interpreted through the word distribution and manually labeled by a word or phrase to reflect the theme and series of activities that intuitively link to the theme. For instance, Topic 6 is associated with "Watching a live show"; Topic 17 reflects "Work activities"; Topic 23 is most likely about "Meals"; and Topic 26 is about "Education". This addresses the first part of RO1 that the derived topics are semantically meaningful and associated with certain activities.

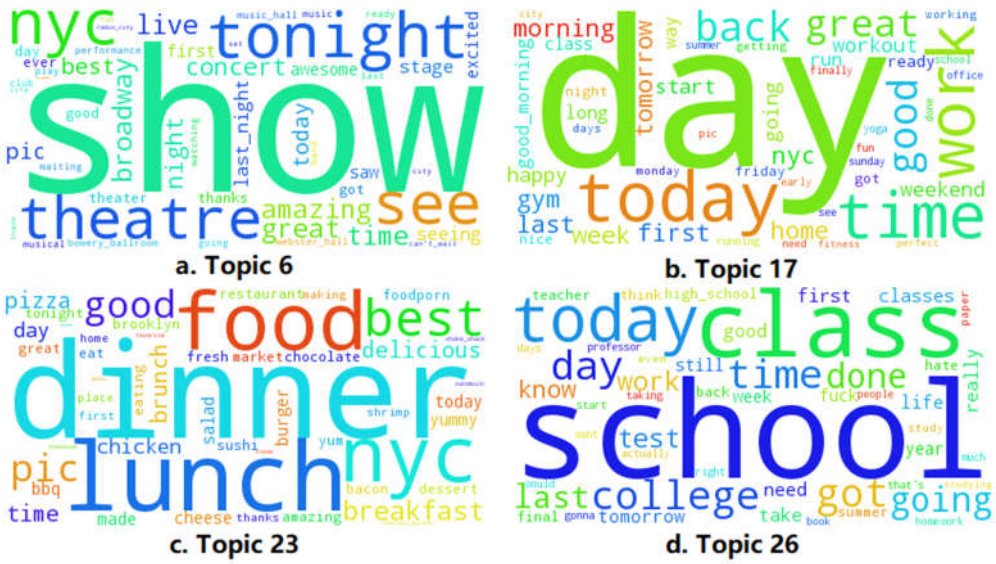


Figure 3-2 Word-cloud of sample topics. A larger font size indicates a higher word probability. Note that the word weights have been normalized to allow comparison between topics.

4.1.4. Perplexity and Number of Topics

The number of topics (T) in a set of documents is an important priori parameter for an LDA model. If a small number of topics is selected, there might be a risk that the discovered topics might be too coarse and that one topic may cover a variety of activities that are different by nature. On the other hand, a large number of topics may

lead to a difficulty in interpretation. One metric to evaluate the "goodness" of an LDA output model and determine the best number of topics is *perplexity* as suggested by (Blei et al., 2003). Perplexity, in this case, is defined as:

$$perplexity(D) = e^{\frac{-\sum_{d=1}^{M} logp\,(w_d)}{\sum_{d=1}^{M} N_d}}$$

where D is a set of test documents that are held from the document set for building the LDA model; M is the size of D; N_d is number of words in a document d from document set D; and $p(w_d)$ is the probability of word distribution in the document. A lower perplexity indicates that the output of the probabilistic model is "better," though a larger number of topics have a lower perplexity generally. The trade-off, however, is that an LDA model with a lower perplexity can be less meaningful with respect to semantic interpretation as reported by Chang et al. (2009). Zhao et al. (2015) suggest a heuristic approach to balance the issue by using an additional metric, the *rate of perplexity change* (RPC), to determine a proper number of topics. RPC is defined as:

$$RPC(i) = \left|\frac{P_i - P_{i-1}}{t_i - t_{i-1}}\right|$$

where t_i is the number of topics from an increasing sequence of topic numbers; and P_i is its corresponding perplexity. If the condition $RPC(i) < RPC(i+1)$ is satisfied, then the first t_i that matches the condition is the best topic number. Even though there is a trade-off between perplexity-based optimization and semantic interpretation, it has been determined that RPC is a reasonable metric for identifying a reasonable number of topics. By employing the RPC, we used an increment of 10 for the number of topics from 40 to 150, and applied 5-fold cross-validation to calculate the average perplexity

for each topic number to the tweet data set that contains all tweets from the three cities. The change of perplexity is quite small (Figure 3-3). Given this, for this study 90 was selected as a reasonable number of topics that can be distinguished. Manual qualitative evaluation of the resulting topics confirmed this number as well. The tweet number in each topic is highly skewed (Figure 3-4) with a mean 47678.87, and a standard deviation of 54257.48. This is likely due to the uneven intensity of activities. Topics referring to consistent daily activities, such as daily chatting, work, and recreation, have a large tweet number, while event-related activities, such as commenting on a new album, has a low tweet number.

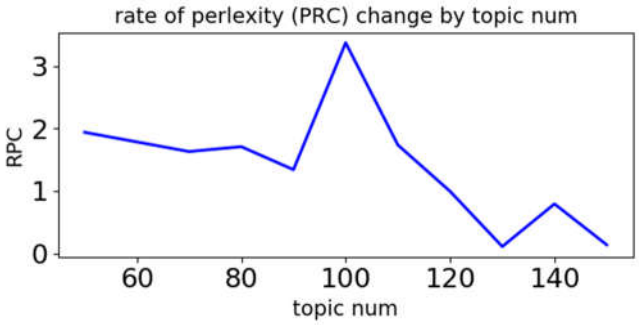


Figure 3-3 RPC for different topic numbers

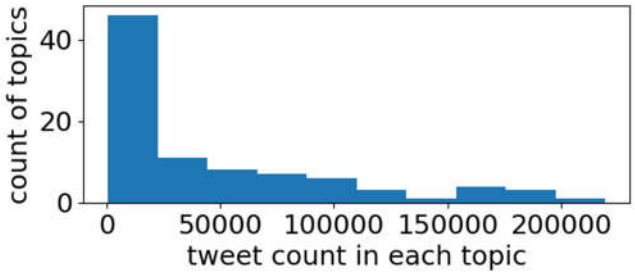


Figure 3-4 Histogram of topic counts by tweets categorized as a topic

4.2. Validate Spatial and Temporal Features of Extracted Activity Topics

The interpretation of the extracted topic word distributions (as of Figure 3-2) shows that ST-LDA can produce meaningful topics that reflect different human activity behaviors. The next step involves validation by ensuring that the extracted topics are associated with certain activity types. This involves three steps that check: 1) if topics have different spatial distribution patterns; 2) if the difference in the spatial distribution is due to the impact or influence of the geographical context, such as land use type or social economic status and 3) if the distribution of each topic over time reflects the attributes of the activity with which the topic is associated.

4.2.1. Temporal Profile of Activity Topics

Figure 3-5 shows the overall aggregated hourly Tweet volume distribution for the three cities by stacking tweets in the same hourly interval on different dates, which can be expressed as:

$$P_{h,c} = \frac{V_{h,c}}{\sum_0^{23} V_{h,c}}$$

where $V_{h,c}$ is the volume of aggregated tweet volume within hth hourly intervals, disregarding the dates. All three profiles have two local peaks around noon and in the evening that we assume to be an overall background temporal signature across all activities.

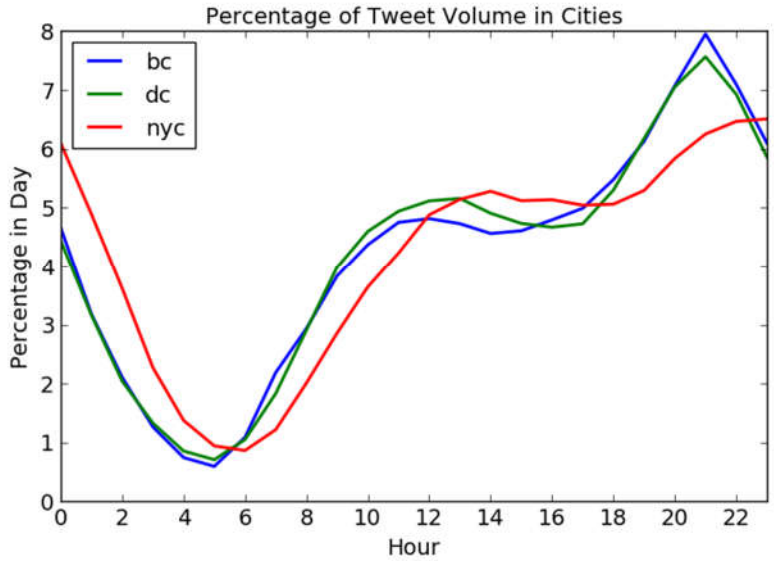


Figure 3-5 Percentage of aggregated Tweet volume by-hour in the three cities.

We further refine the background signature by topic IDs, which are denoted as $P_{t,h,c}$, where t is the topic ID. To characterize how a topic's signature is different from the background signature for a city, the difference ($P_{diff}{}_{t,h,c}$) between a topic and the background is calculated as:

$$P_{diff}{}_{t,h,c} = P_{t,h,c} - P_{h,c}$$

and then the by-hour differences between a topic's temporal profile in a city and the by-hour percentage of the tweet volume in that city can be calculated. If there is no difference between a topic's signature and that of the background, all $P_{diff}{}_{t,h,c}$ values shall be zero. Otherwise, if a $P_{diff}{}_{t,h,c}$ is positive, it means the activity that is presented by a topic is more active than the average topics, and *vice versa* for negative values.

As displayed in Figure 3-6, the temporal profile of the volume of tweets captures and reflects the various activities that are mentioned, as one might expect. In Section 4.1.3, Topic 6 is labeled as "watching a live show" based on the dominant words

extracted from this grouping of subtopics. This topic peaks between 19:00pm and 22:00pm for all three cities (Figure 3-6(a)), while there are fewer tweets on this topic at other times during the day. This peak period appears to reflect what one intuitively expects for a topic related to live entertainment (e.g., theater or concert going). On the other hand, Topic 11 "Work" (Figure 3-6(b)), depicts a very different temporal profile, one that matches commonly accepted "working hours." Activity related to this topic increases above average from approximately 6:00a.m., which coincides with many commuter trips to work in these cities, and remains above average until around 18:00pm. These findings confirm related research on activity and place type temporal patterns for other urban centers, e.g., Ye, et al., 2011 and McKenzie, et al., 2015.

Aside from these more common or expected *local* temporal activity patterns, we also uncover less commonly known *regional* temporal patterns. For example, Figure 3-6(c) shows hourly temporal patterns for the topic "Meals" in our three cities. The temporal profiles in all three cities shows two positive peaks that correspond to lunch and dinner time and that fit with our existing understanding of meals. Interestingly though, these data show that the peak meal times for NYC are approximately two hours later than the peaks for BC and DC. A similar temporal offset is also observed in Topic 26 "education" where the three cities appear to have different peak hours (Figure 3-6(d)). As this topic includes the activities of high school and college students (as demonstrated by the word cloud in Figure 3-2(d)), this offset might be influenced by the different proportions of these two populations of students who attend class according to different schedules. This offset as well as other unique patterns found

between topic signatures, suggests that each city has its own unique temporal activity pulse.

It should be noted that these results rely on data collected over months and aggregated into a single day, reflecting an *average* day profile without considering larger temporal variances, such as the difference between weekdays and weekends, seasonal trends, and any potential influences from large events. However, as being validated, the observations do match our existing knowledge on temporal profile of certain activities as proposed in RO1. There are no controversial observations in the results.

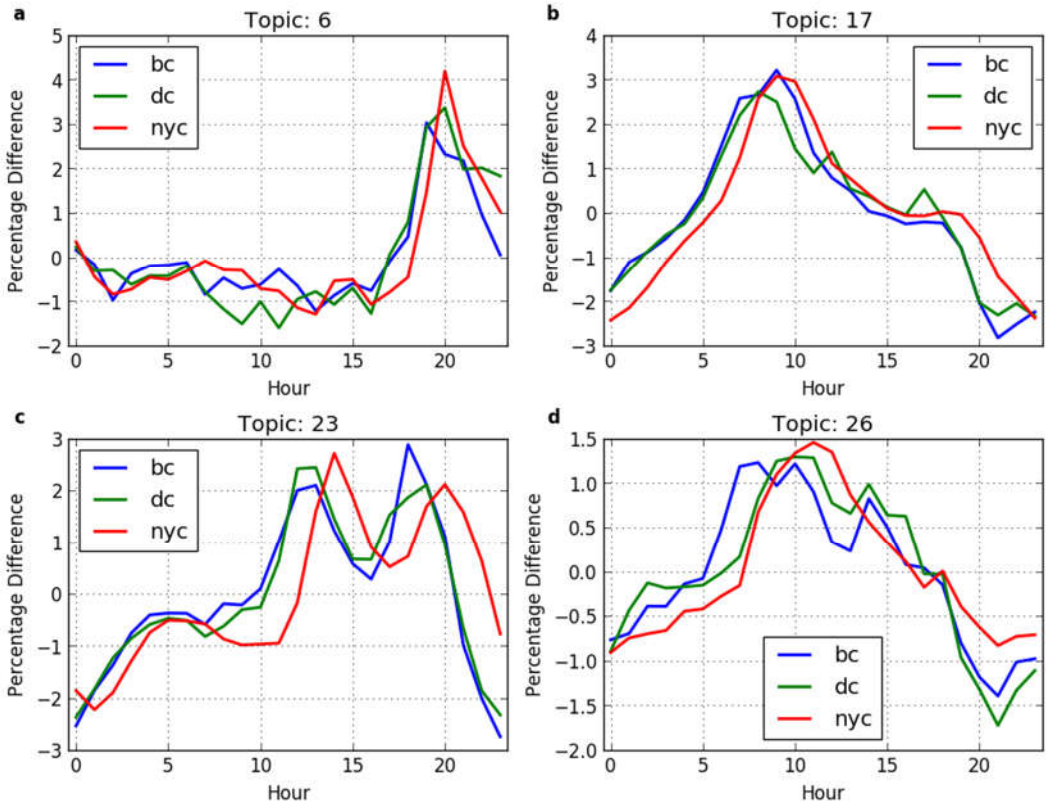


Figure 3-6 Temporal profile of per-hour percentage for selected topics. IDs are corresponding to the word-clouds in Figure 3-2. (a) is suggested as "Watching live show"; (b) is suggested as "Work"; (c) is suggested as "Meals"; and (d) is suggested as "Education".

4.2.2. Spatial Signatures of Topics

The spatial distribution of twitter-based activity topics is analyzed in two ways: within a city and between cities. Many cities have a uniqueness to their activity space, e.g., NYC is very much a city of commerce, whereas Washington D.C. is more politically focused. For this reason, it was expected that the activity topics extracted from georeferenced tweets may differ between cities as was the case with some of the temporal patterns. In addition, the spatial distribution of activity topics within each city is explored.

4.2.3. Local Activity Topics

Given that there are some known differences between the cities in this study, the extracted activities topics were analyzed to determine a set that are unique to one city and those that are not as prominent or not found at all in another city. The ratio of a topic's volume in a city and the topic's expected volume in the city is used to characterize this phenomenon. If a topic is common in all three cities, the volume in a city is expected to be proportional to the total volume of tweets in the city. The ratio ($R_{i,j}$) can be defined as:

$$R_{i,j} = \frac{C_{i,j} - EC_{i,j}}{EC_{i,j}}, \; EC_{i,j} = \frac{C_i}{C_{All}} * C_j$$

where $i \in \{0,1 \dots 89\}$, $j \in \{BC, DC, NYC\}$, $C_{i,j}$ is the volume of tweets labeled as topic i in a city j. $EC_{i,j}$ is the expected volume of tweets labeled as topic i in a city j. C_i is the volume of tweets labeled as topic i in all three cities, whileC_j is the volume of tweets in a city j. C_{All} is the total volume of tweets in all three cities. As demonstrated by

Figure 3-7, there are several topics that are highly localized. For example, Topic 2 exists at a level of 50% more than is expected in NYC. Many of the top words are associated with NYC toponyms (Figure 3-8(a)). Topic 18, on the other hand, is more than 3 times higher than expected in DC, where the top words are associated with *policy* (Figure 3-8(b)). In contrast, topics that are associated with common activities as discussed previously (e.g., Topics 6, 17, 23, and 26), have very low offsets from zero, implying that they are proportionally distributed across each city's total tweet volume.

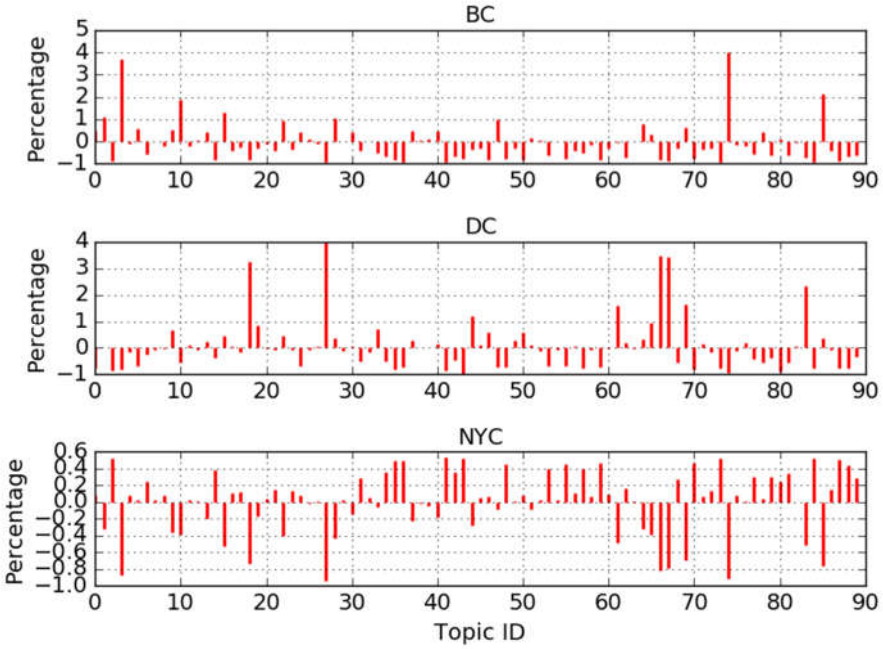
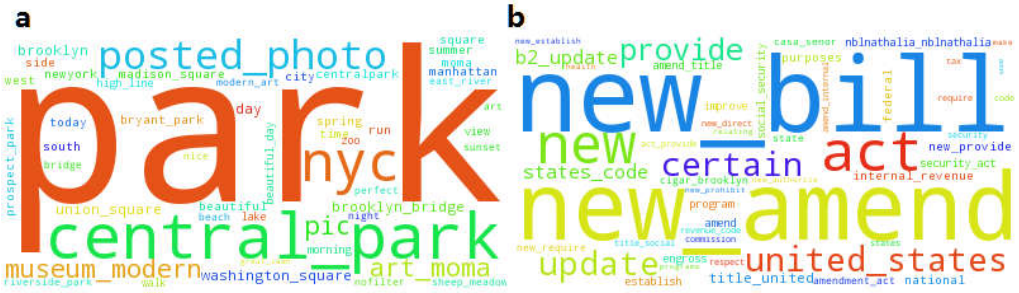


Figure 3-7 $R_{i,j}$ for topics in each city



(a) (b)

Figure 3-8 Word-cloud of (a)Topic 2 and (b) Topic 18

4.2.4. Mapping Activity Topic Patterns

Mapping the density of an activity topic in a city helps to understand the spatial distribution of the topic and its associated activity. The hotspots of a topic are identified where high-density areas cluster on a map, and the geographical context of a cluster on local maps as well as land use maps are checked to explore if the spatial clusters are associated with a certain geographical context, such as the type of land use. Figure 3-9 - Figure 3-12 show the spatial distribution of selected activity topics in the three cities by using Kernel density with 100-meter grids. The results show that there is a strong correspondence between the hotspots of topic clusters and the function of the places in which the clusters are located. For example, the two clusters of Topic 18 (Politics) highlight two of the most important political sites in DC, namely the *White House* region and the area around the *Capitol Building* (Figure 3-9). Similarly, Topic 23 (Meals) are clustered in restaurant-dense regions of BC (Figure 3-10) and DC (Figure 3-11). Topic 26 (Education) identifies numerous educational institutions in NYC including middle schools, high schools, and colleges (Figure 3-12). The spatial distribution of the topics shows the property of the activities as proposed in RO1. In addition, it also confirms part of our claim in RO2 that the derived activities can profile the neighborhoods in a city as a signature of how people interact with the urban space. Combining the spatial and temporal profiles of the derived topics, the first null hypothesis is rejected. The selected samples have demonstrated that even if the topics are modeled combining the spatial and temporal profiles of the derived topics, the first

null hypothesis is rejected. The selected samples demonstrate that even if the topics are modeled semantically, their distributions in space and time are unique.

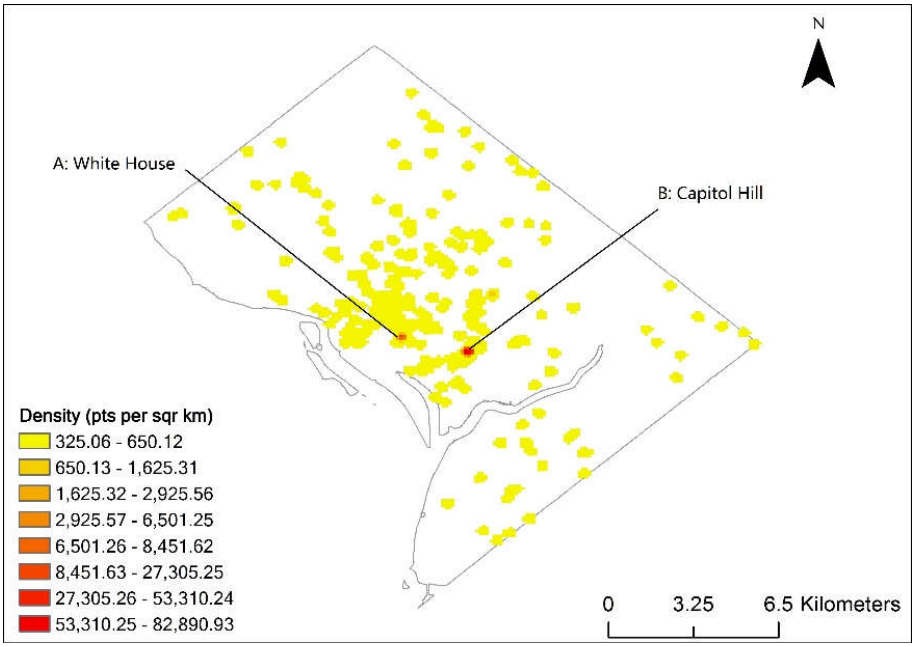


Figure 3-9 The geography of Topic 18 "Political" in DC. Place A: the White House. Place B: the Capitol Hill. (Tweets in water-body and parks are masked out)

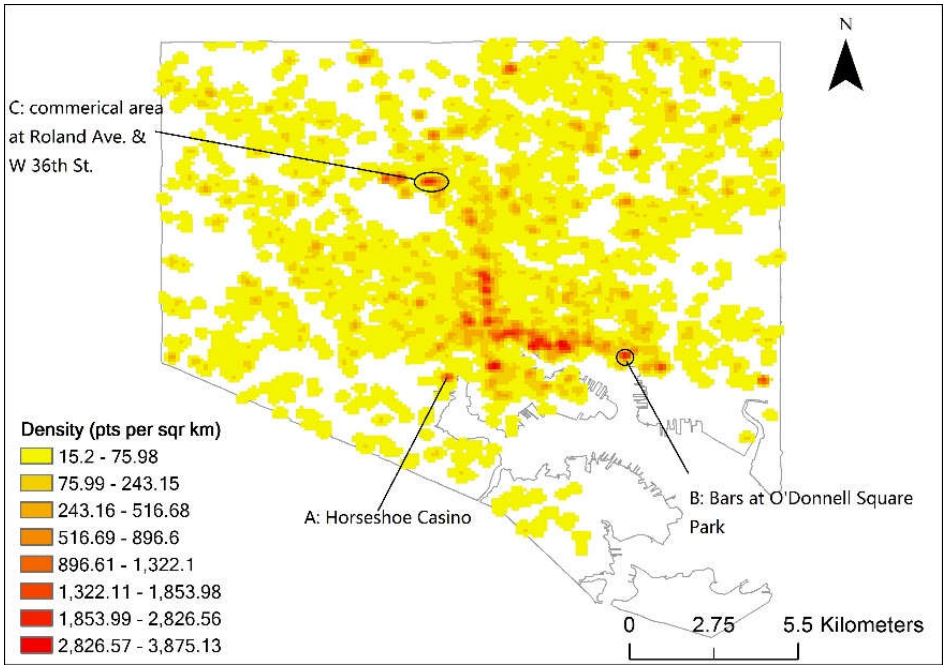


Figure 3-10 The geography of Topic 23 "Meal" in Baltimore City. Place A: the Horseshoe Casino. Place B: a bar area near the O'Donnell Square Park. Place C: a commercial area with bars and restaurants around the intersection at Roland Ave. and W 36th St.

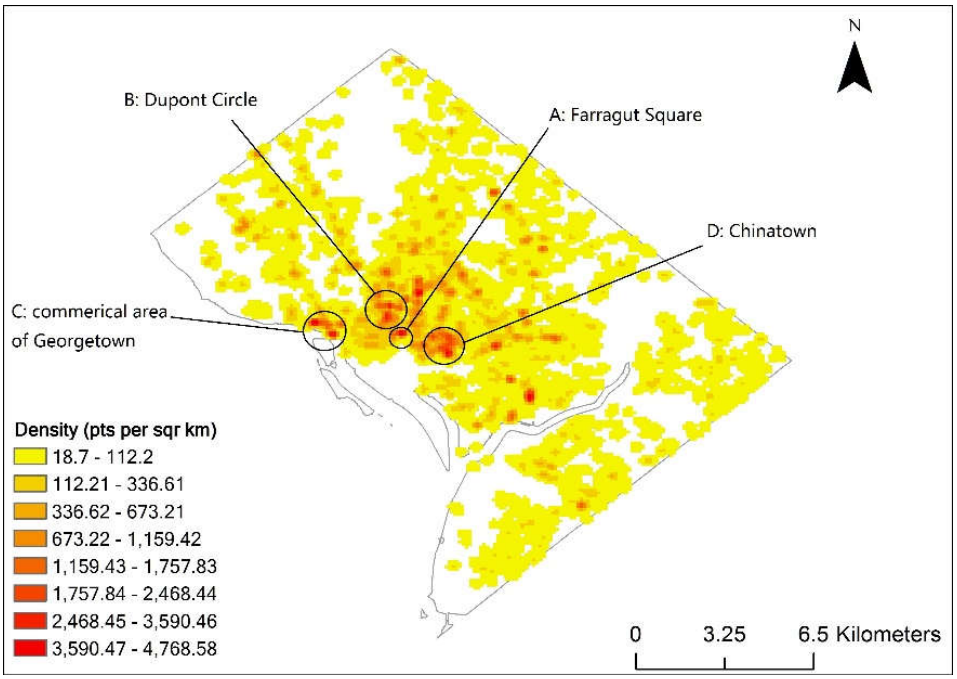


Figure 3-11 The geography of Topic 23 "Meal" in DC. Place A: the commercial area at Farragut Square. Place B: the commercial area around the Dupont Circle. Place C: the commercial area at Georgetown. D: Chinatown in DC.

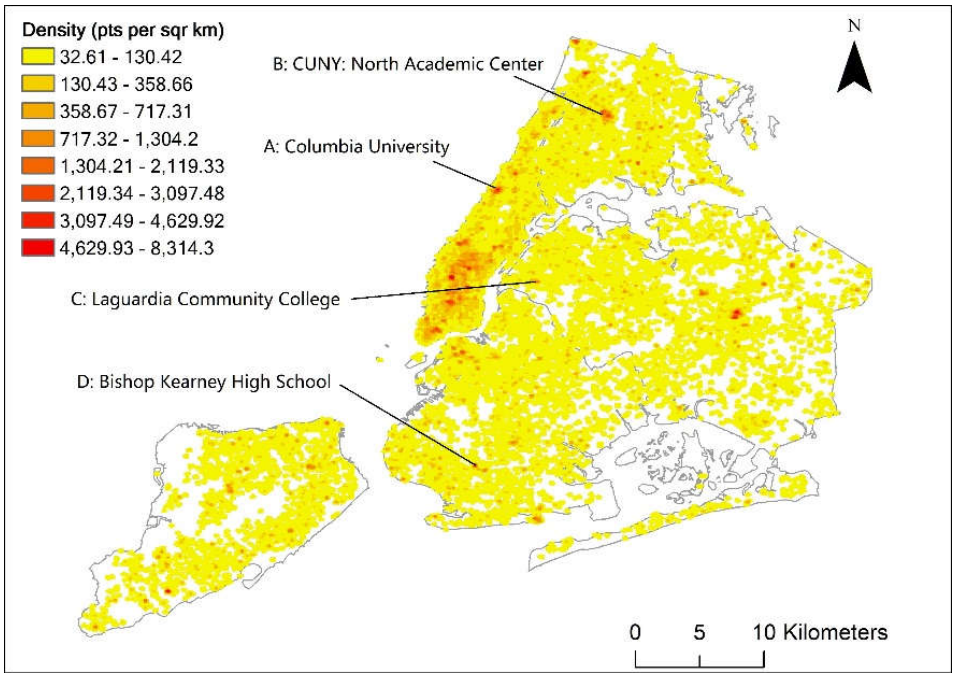


Figure 3-12 The geography of Topic 26 "Educational" in New York City. Place A: Columbia University. Place B: the North Academic Center of the City University of New York (CUNY). Place C: LaGuardia Community College. Place D: Bishop Kearney High School

5. *Activity Patterns at the Neighborhood Level: Similarity Within and Between Cities*

The previous sections have demonstrated that temporal and spatial patterns of activity topics capture the range of human activity behavior within a city. This work also shows that activity topics are not evenly distributed in time and space. The distribution pattern is strongly associated with the nature of the activities that are represented by a topic. This finding indicates that the distribution of topics representing activities can be used to differentiate neighborhoods within a city. Following on this finding, we designed an approach to find similar neighborhoods within a city and between cities based on an activity. The distribution of extracted activity topics is used to represent each neighborhood. This may be helpful for people who move to a new city but want to maintain their life style by living in a neighborhood that is similar to their neighborhood in their original city. To accomplish this task, the neighborhood boundary from each city's planning department is used as the geographical units on which to aggregate topic distributions. While neighborhood boundaries are often fuzzy in nature, they are typically defined based on socio-demographic characteristics, land-use, and urban planning designations. The names of the neighborhood are often selected by the local planning department and adopted for use by the residents of the neighborhood, which is helpful to guide and validate our results, though people may not have the exact sense of a place in terms of boundary in geography. There are 278 neighborhoods in BC, 126 neighborhoods in DC, and 195 neighborhoods in NYC. Topic distributions are determined for each neighborhood by summing and normalizing

the activity topics by topic ID. From this, a vector with 90 items for each neighborhood, whose *i*th item is the number of tweets labeled with topic ID *i*. This vector set is identified as *ALL_TOPIC.* Considering that some topics are local to a specific city as discussed above, a subset of topics are excluded from analysis. Any topic with an absolute value of $R_{i,j}$ larger than 0.5 for any city is excluded from the subset. After the filtering, 35 topics are kept and identified as *COMMON_TOPIC*.

5.1. Measuring Similarity

To measure the similarity between neighborhoods (topic vectors), two distance measures are used: *cosine distance,* which is a commonly used similarity measure in information science (Sankaranarayanan, et al., 2009; Fu, Samet, & Sankaranayananan, 2014) and *Jensen-Shannon divergence* (JSD*, Lin, 1991).* Cosine distance is the measure of the angle of two vectors and is defined as:

$$cosine_{distance}\ (A, B) = 1 - \frac{A \cdot B}{|A| \cdot |B|}$$

where A and B are two vectors, and $|\cdot|$ is the norm of a vector.

JSD measures the similarity between two probability distributions, which is a symmetrized and smoothed version of the Kubell-Leibler divergence. JSD is defined as:

$$\mathrm{JSD}(P||Q) = \frac{1}{2} D_{KL}(P||Q) + \frac{1}{2} D_{KL}(Q||P)$$

$$D_{KL}(P||Q) = \sum_i P(i) log \frac{P(i)}{Q(i)}$$

where P and Q are two probability distribution; $P(i)$ is *i*th item of P.

After computing the two metrics, these were then applied to calculate pair-wise distances of any two neighborhoods respectively, resulting in two similarity matrixes.

5.2. Similarity Matrix Visualization

After calculating the pair-wise distance between two neighborhoods, multidimensional scaling (MDS, Kruskal, 1964) is employed to visualize the distance matrix as this metric can reduce the dimension to 2 while preserving the inter-object distance. MDS uses *Stress* that ranges from 0 to 1 to measure the goodness of MDS, and where 0 represents a good fit. The smaller distance between two data points in a MDS figure shows more similarity.

The visualization results show the influence of inter-city characteristics as a slight clustering effect for each city in the plots using COS (Figure 3-13a) and also JSD (Figure 3-13b) can be observed. For a given neighborhood in a city, it can also be observed that there are always some neighborhoods in the other cities that may be closer in similarity than neighborhoods within the same city. After removing the local topics, the results show the distinctions between the three cities tend to disappear and the distributions using both COS (Figure 3-13c) and JSD (Figure 3-13d) mostly overlap. All MDS have a fair goodness of fit where Figure 3-13a and Figure 3-13b have a stress

value 0.19 and 0.17, respectively; Figure 3-13c and Figure 3-13d have a stress value 0.16 and 0.17, respectively.

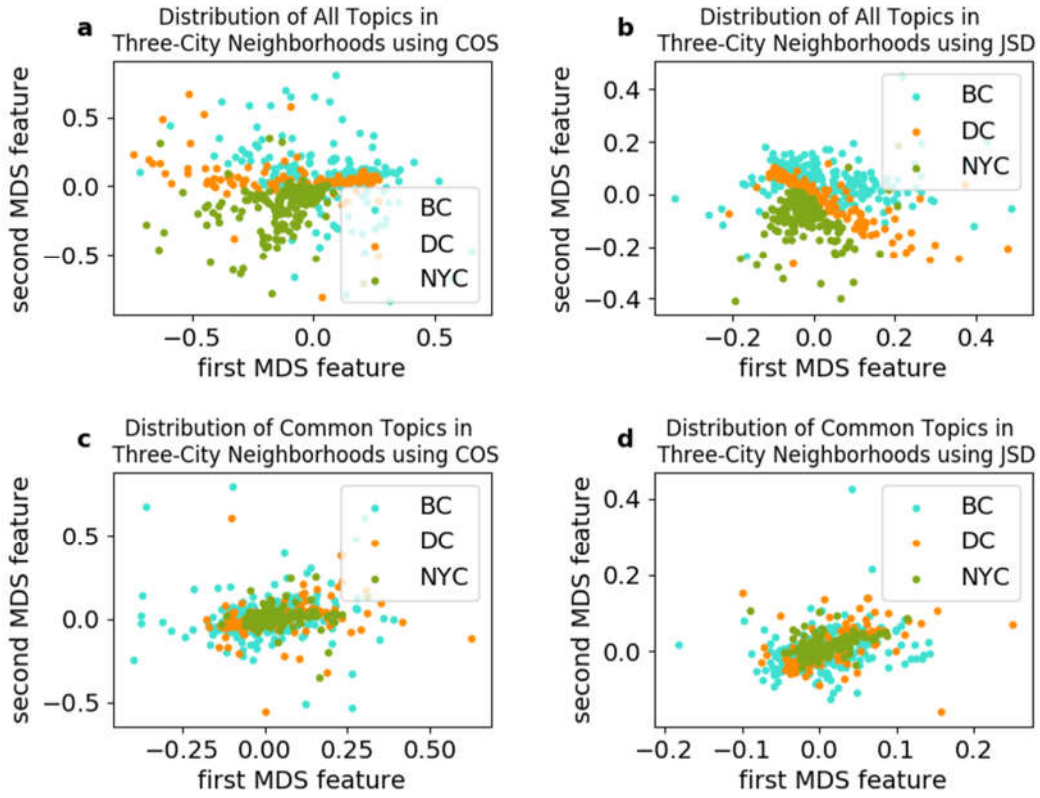


Figure 3-13 MDS of (a) cosine-distance based ALL_TOPIC distribution of neighborhoods, (b) JSD-based ALL_TOPIC distribution of neighborhoods, (c) cosine-distance based COMMON_TOPIC distribution of neighborhoods, and (d) JSD-based COMMON_TOPIC distribution of neighborhoods.

6. *A Case Study*

A qualitative method is initially employed to validate if the activity-based similarity can return meaningful results. Similar neighborhoods in the same city, between the other two cities, and in all three cities are explored based on the two metrics and two topic sets. For demonstration, the sample results of the neighborhood *Dupont Circle* is displayed in Table 3-1. Dupont Circle is a historic district in the northwest of Washington D.C. The neighborhood has a diverse geographical context, including a

traffic circle, park, farmers market, embassies, and restaurants. The top topics in this neighborhood are those identified as “party” (Topic 16), “dinner” (Topic 23), etc. As Table 3-1 shows, the corresponding results discovered via cosine distance and JSD have high agreement. The geography of neighborhoods in the three cities and their cosine distance to Dupont Circle can be compared (Figure 3-14). For the top similar neighborhoods inside DC, we find that neighborhoods Shaw, U Street Corridor, and Adams Morgan are most similar. Consequently, these neighborhoods are also directly adjacent to Dupont Circle, showing the influence of Tobler’s’ First Law (Tobler, 1970). Chinatown in DC is not spatially close to the Dupont Circle neighborhood, but it also has numerous restaurants that demonstrate similar social functions and activity affordances. Such similarities can be found in the most similar neighborhoods in the other two cities. For example, the East Village in New York City also contains a diverse culture and historically, it has experienced gentrification similar to that of Dupont Circle. Similarity can also be observed by comparing the result of ALL_TOPIC and COMMON_TOPIC that suggests that neighborhoods in DC are consistent, while the recommended, most similar neighborhoods in NYC and BC are slightly different. One could also observe that the neighborhoods within DC, i.e., the same city, are more similar to Dupont Circle when taking into account of all the topics. However, it is difficult to determine which topic set actually models the similarity between the neighborhoods better, since the two sets may characterize the nature of activities from different perspectives.

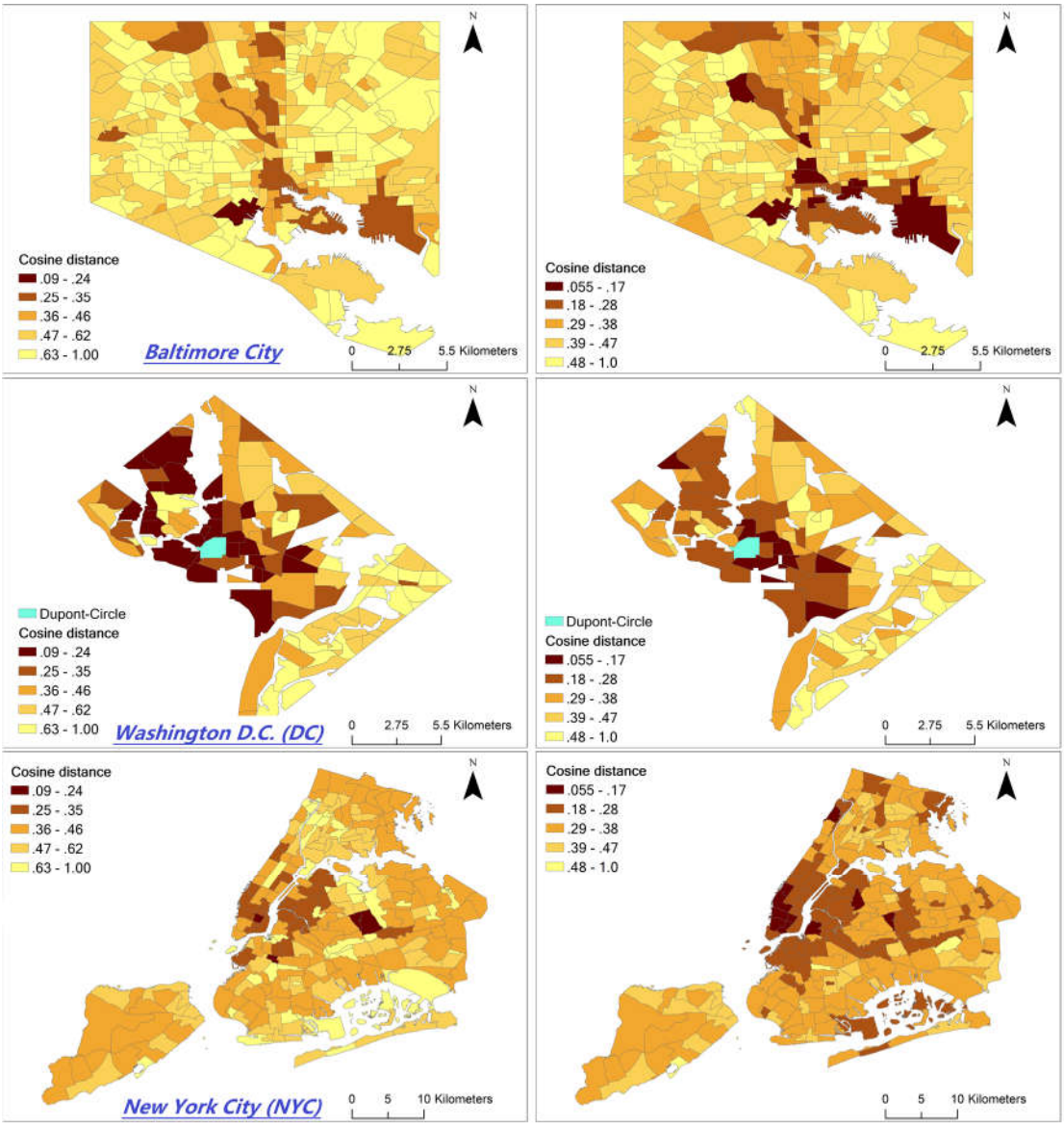


Figure 3-14 Cosine distance between Dupont Circle and the other neighborhoods in BC, DC and NYC. Legends in BC and NYC are coordinated to the legend of DC.

Table 3-1 Top 5 most similar neighborhoods in different cities for neighborhood "Dupont Circle" in DC with different topic sets and different similarity metrics. Distance value is displaced under the neighborhood name.

ALL_TOPIC							
COS				**JSD**			
BC	**DC**	**NYC**	**THREE**	**BC**	**DC**	**NYC**	**THREE**
Carroll - Camden Industrial Area 0.21	U Street Corridor 0.09	Prospect Heights 0.23	U Street Corridor (DC) 0.09	Carroll - Camden Industrial Area 0.11	U Street Corridor 0.05	Prospect Heights 0.12	U Street Corridor (DC) 0.05
Charles Village 0.26	Mount Vernon Square 0.12	Rego Park 0.24	Mount Vernon Square (DC) 0.12	Downton West 0.12	Adams Morgan 0.05	Gramercy 0.12	Adams Morgan (DC) 0.05
Mount Vernon 0.26	Adams Morgan 0.13	East Village 0.24	Adams Morgan (DC) 0.13	Inner Harbor 0.12	Mount Vernon Square 0.06	Murray Hill-Kips Bay 0.12	Mount Vernon Square (DC) 0.05
Cedarcroft 0.26	Atlas District 0.14	Greenpoint 0.26	Atlas District (DC) 0.14	Tuscany-Canterbury 0.12	Atlas District 0.06	East Village 0.12	Atlas District (DC) 0.06
Johns Hopkins Homewood 0.27	Shaw 0.14	Murray Hill-Kips Bay 0.26	Shaw (DC) 0.14	Charles Village 0.13	Chevy Chase 0.07	West Village 0.12	Chevy Chase (DC) 0.07
COMMON_TOPIC							
COS				**JSD**			
BC	**DC**	**NYC**	**THREE**	**BC**	**DC**	**NYC**	**THREE**
Downtown 0.08	Chinatown 0.05	Chinatown 0.08	Chinatown (DC) 0.05	Downtown 0.03	Chinatown 0.02	Chinatown 0.03	Chinatown (DC) 0.02
Federal Hill 0.09	U Street Corridor 0.06	Clinton 0.09	U Street Corridor (DC) 0.06	Federal Hill 0.04	U Street Corridor 0.02	Clinton 0.03	U Street Corridor (DC) 0.02
Fells Point 0.12	Shaw 0.07	Hudson Yards-Chelsea-Flatiron-Union Square 0.10	Shaw (DC) 0.07	Fells Point 0.04	Downtown 0.03	Hudson Yards-Chelsea-Flatiron-Union Square 0.03	Chinatown (NYC) 0.03
Carroll - Camden Industrial Area	Atlas District 0.07	Rego Park 0.10	Atlas District (DC)	Upper Fells Point 0.05	Atlas District 0.03	North Side-South Side 0.04	Downtown (DC) 0.03

0.13			0.07				
Canton Industrial Area 0.14	Downtown 0.11	North Side-South Side 0.12	Chinatown (NYC) 0.08	Woodberry 0.05	Shaw 0.03	East Village 0.04	Atlas District (DC) 0.03

To determine if the cosine distance and JSD have consistent agreement with respect to similarity for all neighborhoods, mean reciprocal rank (MRR, Voorhees, 1999) and normalized discounted cumulative gain (NDCG) are employed. MRR is defined as:

$$\text{MRR} = \frac{1}{|Q|}\sum_{i=1}^{Q}\frac{1}{rank_i}$$

where Q is a query, and $rank_i$ refers to the rank position of the first relevant item for the *i-th* query in the target list. In this study context, the first suggested neighborhood from one distance is treated as the source query, search its rank in the suggested list from the other distance that serves as the target list, and use the ranks to calculate the MRR for all neighborhoods.

NDCG is used for measuring the ranking quality based on discounted cumulative gain (DCG), which is defined as:

$$DCG_p = \sum_{i=1}^{p}\frac{rel_i}{log_2^{(i+1)}}$$

where p is a rank at a returned list as the target, and rel_i is the weighted relevance of result at position *i*. NDCG then can be computed as:

$$nDCG_p = \frac{DCG_p}{IDCG_p}$$

$$IDCG_p = \sum_{i=1}^{|REL|}\frac{2^{rel_i}-1}{log_2^{(i+1)}}$$

where |REL| is the list of results ordered by their relevance, which is the ideal order.

If NDCG is close to 1, the two orders have good agreement. In this study, we set $rel_i = \frac{1}{i+1}$, and set p to 10. Both metrics were calculated using the Python package "rank_metrics"[6]. The mean NDCG is used to evaluate the overall performance.

As a first step, the neighborhood results that include all neighborhoods in the three cities as candidates are assessed. For the ALL_TOPIC set, using the cosine-distance-based similarity returned neighborhoods as the benchmark, the MRR is 0.56, and using the JSD-based returned neighborhoods as the benchmark, the MRR is also 0.56. For the COMMON_TOPIC set, the values of MRRs are both 0.67. This implies that in both topic profile contexts, the most similar neighborhood returned based on one type of distance can also be found within the top two similar neighborhoods. It also shows that using the COMMON_TOPIC set, agreement is slightly better. This conclusion is further validated by applying a t-test on the pairs of reciprocal rank lists with $p < 0.01$. The NDCG shows similar results for the ALL_TOPIC context with the average NDCG values both equal to 0.70 using either the neighborhood similarity scores computed using cosine-distance based similarity or the JSD-based approach. In the COMMON_TOPIC context, both average NDCG values are 0.80.

Second, the neighborhood results are checked for results that excludes the neighborhoods in the same city. For example, for a neighborhood in BC, what are the most similar neighborhoods in DC and NYC? As Table 3-2 shows, there is strong agreement between the results using two similarity metrics.

6 https://gist.github.com/bwhite/3726239

By exploring the topic distribution in a neighborhood and finding out its similar neighborhoods, we rejected the second null hypothesis because the activities, by using the derived topics as proxy, can characterize a neighborhood.

Table 3-2 MRR and mean NDCG for the neighborhood suggestion that excludes the neighborhoods in the same city. The subscript COS-JSD means using the suggestion from cosine distance-based ranking as benchmark to evaluate the suggestion from JSD based ranking. Vice versa.

	$MRR_{COS-JSD}$	**$MRR_{JSD-COS}$**	**$MEAN\_NDCG_{COS-JSD}$**	**$MEAN\_NDCG_{JSD-COS}$**
	TOTAL_TOPIC			
BC	0.67	0.68	0.81	0.81
DC	0.60	0.58	0.76	0.75
NYC	0.58	0.57	0.78	0.77
	COMMON_TOPIC			
BC	0.74	0.74	0.83	0.83
DC	0.71	0.71	0.86	0.86
NYC	0.66	0.68	0.84	0.85

7. *Conclusions & Future Work*

This study proposed that socially sensed information in georeferenced tweets can be usefully deployed as a proxy to identify activity types in space and time within a city. Topic modeling is applied as a tool to extract activity topics from a massive tweet dataset to reduce complexity and time, and to identify activities. Manual interpretation of the word distribution of the extracted topics, confirms that topic modeling can extract meaningful topics as a proxy for corresponding activity types from datasets with massive amounts of tweets. Further validation of the extracted topics' distributions in time and space showed that the theme of a topic is related to the nature of the activities, as well as to geographical context, such as land use corresponding to where the activities occur. These results demonstrate that deconstructing activities in a city into different activity types using an NLP approach on tweet text, may help to reveal and identify detailed activity patterns in a city. It also

indicates that some tweeting activities, even as behaved in cyberspace, are bounded by geographical context.

This research also showed how these extracted topics can be used as attribute features for profiling social functions of a neighborhood. We selected neighborhoods in three U.S. cities for a case study and validated our analysis based on two different distance metrics. We found that the similarity between neighborhoods based on the activity profiles are consistent. The suggested similar neighborhoods matched information on the neighborhoods with respect to the similarity from a social, economic and historical (e.g., urban development) perspective.

To conclude, we explored an attempt to quantitatively profile a neighborhood from the signatures of human activities referenced by individuals. This provides a new perspective different from demography-based profiling and descriptive profiling. As the study is based on multiple cities, it has the potential to easily extend to more cities and collect knowledge on urban geography in an automated way.

This study demonstrated that by employing NLP methods for analyzing georeferenced text from Twitter, it is possible to build a model that links the posting of activities online to real-world activities. This enables researchers to differentiate more detailed activities in Twitter data beyond simply treating all tweets as identical activities or using key-word based models. Even though the proposed methods helped to discover interpretable activities and their spatial and temporal distributions in cities from the Twitter dataset, it must be noted that georeferenced tweets have a limitation as a perfect unbiased proxy to actual activities. The georeferenced tweets only take about 1% of the overall tweets (Morstatter, et al., 2013), which potentially can be biased

from the population of Tweets. The demography of Twitter users may not be the same as the population's demography, which may lead to differences between the sensed activities from tweets and the real activity distribution. One possible solution for reducing the impact of the bias is to integrate different data sources, e.g. point-of-interest data or other georeferenced documents, that potentially have different biases, and compare the results to generate a more robust and general conclusion. Another solution is to infer the users' demography by combining different survey sources (Lansley & Longley, 2016) and correct the bias by calibration using the demography of the population (Longley & Adnan, 2016). In addition, Twitter users do not post tweets consistently. Users may have very different habits, for example, some users may post a large number of tweets in a day, while others may merely post a single tweet. The interval between two tweets from the same user can also vary greatly. In this study, each tweet is simply treated as a single activity. However, further study is needed to investigate a more sophisticated definition to model an individual continues activity by grouping tweets. Similarly, even though most of the derived topics are easy to associate with some activities, it is hard to determine a clear activity theme for other topics, as the semantics reflect several activities. This can also involve discussion about the ontologies that underlie the definition of a unique *activity*, such as Wang and Stewart (2015) discussed. Therefore, our findings are based on the available data sets and can only be used as reference rather than scientific ground truth to present the population with the awareness of all potential biases.

Chapter 4: : Integrating Remotely Sensed Imagery and Activity-Based Geographic Information to Sense Built-Up Land Use Changes in US Metropolitan Areas

1. *Abstract*

Land use structure is a key component for understanding the complexity of urban systems as it provides insights into how people use space, as well as a snapshot of urban dynamics. This paper integrates socially-sensed activity data with remotely sensed imagery to infer land use in a metropolitan area. The proposed approach integrates an impervious surface cover change product from remote sensing as the physical signature of land use, with activity signatures derived from georeferenced tweets to infer land use that involves conversions from undeveloped to developed usages. A case study is conducted to identify land use change in the Washington D.C.-Baltimore metropolitan area between 1986 and 2008. A classification model utilizing both physical and activity signatures was developed to differentiate residential and non-residential places over time. Model assessment shows that the proposed classification workflow differentiated residential and non-residential uses at an accuracy of over 80%. Using the temporal information from remotely sensed imagery, the study also reconstructs the temporal trajectory of development for different land use types. Results indicate that the proposed approach is useful for mapping detailed land use in an urban region, and serves as a new and viable way forward for land use surveying that could be especially useful for megacities and other massive extents.

2. *Introduction*

The world is rapidly urbanizing. By 2014, 54% of the world's population were living in cities, and 2.5 billion more people were projected to be city dwellers by 2050 (United Nations, 2014). With more people residing in urban and suburban areas, there comes a much higher demand for developed space in cities. By 2030, it is forecasted that the global urban land area may triple from the coverage that existed in 2000 (Seto, Guneralp, & Hutyra, 2012). Information on land use (the social function of land) is important for understanding the dynamics and complexity of urban systems. Specifically, the intra-city land use structure can benefit models of carbon emission estimations (Glaeser & Kahn, 2010; IPCC, 2014), hazard resilience (Burby, et al., 2000), and transportation (Iacono et al., 2008; Waddell et al., 2010).

However, we frequently have limited knowledge about the extent of sprawl (i.e., uncoordinated city growth (Batty, Besussi, and Chin, 2003)) of newly-built developments in urban areas. Official land use maps based on land surveying are often not updated frequently due to financial and time costs, and thus do not capture the rapid changes of urbanization. Remote sensing has been successfully applied to projects involving the mapping of land cover in massive urban areas (e.g., megacities), and has contributed to understanding the sprawl of built-up urban areas (Xian, Homer, and Fry, 2009). For example, Song et al. (2016) provided an annual impervious surface change map that mainly captures changes from undeveloped land to a built-up area for identifying locations of urban sprawl. In urban areas, land cover change is often a result of direct human land use change. As useful as it is, remote sensing imagery has a major limitation when it comes to inferring land use, however, and that is due to that fact that

satellite imagery can provide only the physical properties of the surface (Herold et al., 2005), and not necessarily the actual use of land, especially of buildings in an urban context, that is tied more directly to the purposes and activities that individuals associate with these structures.

Recently, socially-sensed geographical data (Y. Liu et al., 2015) that capture human activities on a massive scale have been introduced to model the land use of parcels or the function of places in cities, through applying call detailed records (CDRs, Pei et al., 2014; Reades, Calabrese, and Ratti, 2009; Soto and Frias-Martinez, 2011), georeferenced tweets (Crooks et al., 2015; Frias-Martinez, Soto, et al., 2012; Lee et al., 2012), taxi trajectories (Guo, et al., 2012; Yuan, Zheng, and Xie, 2012), wireless data requests (Nishi, Tsubouchi, and Shimosaka, 2014) and photos from Google Street View (Li, Zhang, and Li, 2017). These data, referred collectively as *socially-sensed data* are used as a proxy for activities in space and time. Usually, the data are first aggregated based on some specific geographic unit (e.g., land parcels or grids), then the data's variances over time are modeled as signatures of the activities (Zhou and Zhang, 2016).

Besides the temporal variances of socially-sensed data, georeferenced text provides additional information on activities. Latent Dirichlet Allocation (LDA, Blei, Ng, and Jordan, 2003) models used in natural language processing (NLP) assume that the observed documents, as a set of words, are associated with a set of unobserved latent topics. A topic is presented as a unique word probability distribution. The process of an LDA model is designed to discover the latent topics and assign these topics to documents. LDA and its variants thus are commonly used for summarizing and

classifying georeferenced documents to find geographical meaning and use of places (Hu and Ester, 2013; Crooks et al., 2015; McKenzie, Adams, and Janowicz, 2015).

Socially-sensed data have their own limitations, however. Most socially-sensed data, e.g. CDRs, taxi trajectories, and georeferenced tweets, are point-based, and do not cover the whole space seamlessly. Therefore, utilizing socially-sensed data usually relies on pre-defined geographic units for aggregating data as most previous studies have done. Similar to issues with remotely sensed data, these pre-defined geographic units, such as road-segmented parcels or zoning parcels, are not always updated frequently, and thus may be outdated. In addition, most socially sensed data sources are held by private companies, and require a study-by-study license to access the data. Another limitation of socially sensed data is the lack of a historical archive due to the fact that the data relies highly on the prevalence of GPS-embedded devices, especially smartphones that have only become widely available in the past decade. Therefore, it may be difficult to model the process of how different types of land uses have expanded over time. As most socially sensed data are collected by GPS-embedded devices, the location accuracy is subject to the device and the environment context (e.g., open space and in-door). Lastly, and perhaps more importantly, the demographic bias (Duggan, 2015) in socially sensed data may limit their applications to certain types of activities and population groups, lacking generalization.

In this study, we propose an integrated framework that uses both socially-sensed data and remotely sensed imagery to characterize land use change in an urban area following the general approach of ‘socializing the pixel’ and ‘pixelizing the social’ (Geoghegan et al., 1998). There have been some attempts to integrate these two data

types in social studies. Socially-sensed data were first employed as a source for validating land cover maps (Fonte, et al., 2015) or a cue for narrowing down the study area for remote sensing analysis (Cervone et al., 2016). They are also used to identify the frontiers of urban sprawl (Rodriguez Lopez, Heider, and Scheffran, 2017). There also have been attempts to integrate remote sensed imagery and social sensed CDRs on land use identification (Jia et al., 2018). In this study, we focus on utilizing both data types as the physical signatures, i.e. physical properties of a land parcel, and behavioral signature, i.e. properties derived from activities on a land parcel, on urban places to identify and differentiate residential and non-residential areas composed of developed land.

This research consists two main research objectives: First, to combine both remotely sensed imagery and socially-sensed human activities data to identify current land uses of areas that have been converted from undeveloped land to built-up land. Second, to estimate the geographic pattern of sprawl for different built-up land uses, i.e. residential and non-residential uses, arising from the result of the first research objective.

The rest of this paper is organized as follows: Section 2 introduces the study area and data collected for the study. Section 3 describes the main workflow that identifies the land use of places in the study area by combining a remote sensing product and socially-sensed activity data. Section 4 demonstrates the main results of the proposed workflow. Section 5 analyzes sprawl in the study area based on the results of Section 4. Section 6 discusses the advantages and remaining issues of the workflow. Section 7 concludes with the main contributions of the paper and proposes future work.

3. *Study Area and Data*

The Washington D.C.-Baltimore metropolitan area was selected as the study area, including the District of Columbia, four municipalities/counties in Virginia, and 17 counties in Maryland (Figure 4-1). The region is the capital of the United States and has experienced rapid urban sprawl between 1984 to 2008. Therefore, this region serves as a strong driver for a study on mapping land cover and land use (Goetz et al., 2003; Sexton et al., 2013; Song et al., 2016).

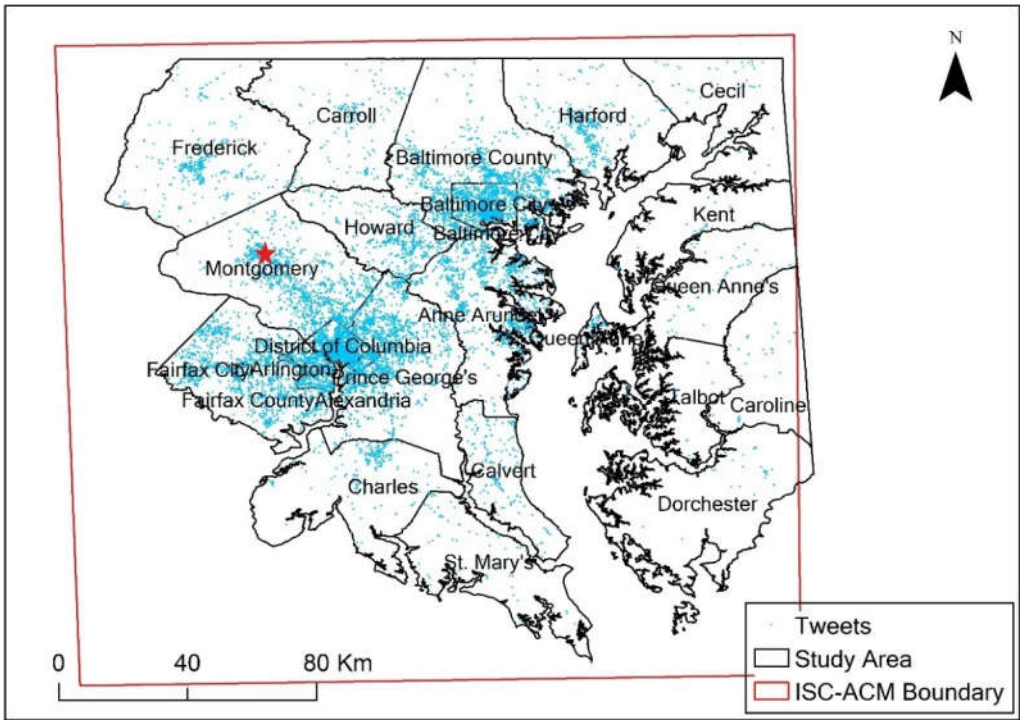


Figure 4-1 Geography of the study area. ISC-ACM stands for Impervious Surface Cover Annual Change Map. Tweets were collected from October 2014 to April 2015. Red star is Rockville, MD with details discussed in Figure 4-2.

Two main data sources were used to map land use and land use change for this analysis. The first source is the Impervious Surface Cover Annual Change Map (ISC-ACM) from the Global Land Cover Facility (GLCF) at University of Maryland (Song et al., 2016). The second data source are georeferenced tweets from Twitter that are

widely used for modeling human activities (e.g. Jenkins, et al., 2016; Hong, et al., 2017). Additionally, land use maps from urban planning departments are employed as a reference to current land use.

The ISC-ACM is a 30-m spatial resolution raster map that identifies land cover change in the Washington D.C.-Baltimore metropolitan area currently covering the period from 1986 to 2008 based on Landsat imagery. Impervious surface cover (ISC) characterizes each 30-m × 30-m pixel as percentage of land surface that cannot be penetrated by water, i.e., paved roads or buildings. The ISC-ACM is composed of three urban growth layers: growth magnitude, growth duration, and growth year (Figure 4-2b, c, d). In a land use/land cover map from local planning department (Figure 4-2a), only the current status of the land is recorded. In the ISC-ACM Change Year layer (Figure 4-2b), each pixel is labeled by the year in which there was a significant increase in the magnitude of impervious surface cover, meaning the land started to change from undeveloped to some degree of being built-up in that year, with ± 1 year uncertainty. The ISC-ACM Change Duration layer (Figure 4-2c) maps the duration of any ISC increase. In the ISC-ACM Change Magnitude layer (Figure 4-2d), each pixel value is the percentage increase of ISC. 80% of all changes are completed with a less than 3-year duration, as Song et al. (2016) finds. Since the ISC-ACM is pixel-based, we can aggregate the pixels using an object-based image processing approach (Blaschke, 2010; Hussain, et al., 2013; Walter, 2004) to join adjacent pixels that belong to the same place as a single object. These objects in turn can be used as the geographic units for aggregating socially-sensed data.

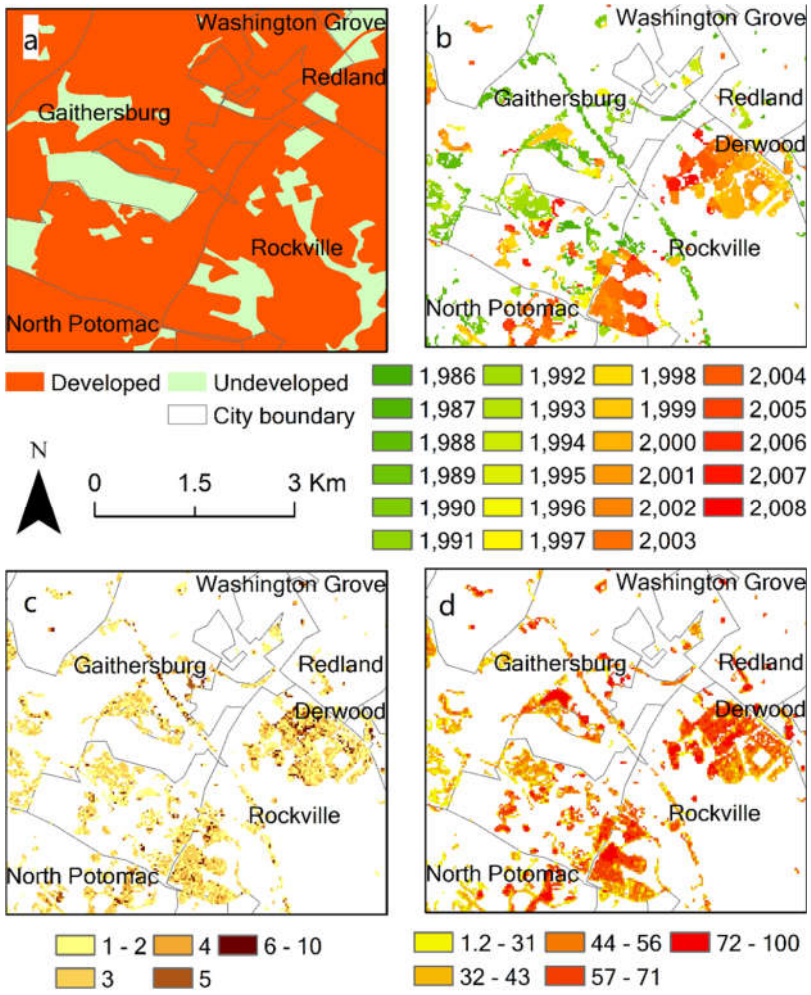


Figure 4-2 North of Rockville, MD (marked as star in Figure 4-1) on the ISC-ACM set layers: a. land use recategorized from the 2010 Maryland Land Use Land Cover Map. b. Change Year layer (time of impervious surface increase), c. Change Duration layer (duration of impervious surface increase in terms of year), d. Change Magnitude layer (percentage of impervious surface increase)

The second major data source, the georeferenced tweets are freely accessible by passing parameters to the Twitter Public Streaming Application Program Interfaces (APIs) with no additional data license required. Given a small enough region, almost all georeferenced tweets can be retrieved (Morstatter et al. 2013). Data were collected from October 2014 to April 2015 via the API. The final data set has ~11.12 million records. There is no information about the position error of tweets. As a reference, the

median horizontal position error of smartphone is reported between 5.0m and 8.5m (Zandbergen and Barbeau 2011).

For an official reference to the current land use in the study area, we utilize the available zoning map or land use map from planning departments that are closest to 2008: For counties in Maryland, this is the 2010 Maryland Land Use Land Cover Map (Maryland Department of Planning, 2010); For Washington D.C., it is the 2006 Land Use Map (DC Office of Planning, 2006); For counties in Virginia, maps are the 2015 zoning maps from each county (Arlington County, 2015; City of Alexandria, 2015; City of Falls Church, 2015; Fairfax County GIS & Mapping Service Branch, 2015). All the detailed land use types are re-categorized into two major land uses: *undeveloped* and *developed*. The undeveloped land uses include forest, water, pasture, cropland, and other natural lands., and the developed land uses include two exclusive sub-types: *residential* and *non-residential*. The non-residential uses include commercial, educational, hospital, industrial, etc.

It should be noted that there are temporal differences among the three types of data sources employed in the study. Because of these differences, we assume that the land use of the regions identified in the ISC-ACM did not change from 2008 to 2015 (the dates of collected Twitter data). In addition, the official land use maps are not frequently updated, and may not reflect the current land cover and land use. For this reason, the official land use maps are used as a reference in the proposed workflow. It is additionally assumed that the built environment land use types in the official maps, such as commercial and residential uses, are correct, while the undeveloped land use types, such as crop land and pasture, may be falsely labeled. These two assumptions

are reasonable because the built-up land uses are unlikely to have been converted back to undeveloped land due to zoning policies and financial costs.

4. Methodology

To identify land use changes, the new developed pixels in the ISC-ACM are grouped first into parcels as the basic geographic unit. Parcels are associated with the official land use maps as a basis for determining both a training set and an unlabeled set. Then, georeferenced tweets are associated with all the parcels. For each parcel, a set of physical properties are calculated as the physical signature and a set of activity properties are derived from associated tweets as the activity signature. Classification models are trained using the training set and then applied to determine the land use type of the unlabeled parcels.

4.1. Deriving ISC Objects

To follow the object-based image processing approach, connected component segmentation (Haralick & Shapiro, 1985) was applied to group adjacent pixels into *objects*. An object can be treated as a place or an *area-of-interest* (AOI, Hu et al., 2015) such as a plaza or a residential community occupying several pixels in the satellite image. It was also assumed that construction of AOIs were continuous in time and space and thus adjacent pixels belonging to the same AOI should be labeled as the same year or adjacent years in the ISC-ACM Change Year layer. Due to the ± 1 year uncertainty of the ISC-ACM (Song et al. 2016), a two-year search radius was designed for the implementation of connected component segmentation in an image processing package Orfeo (Inglada & Christophe, 2009). That is, if the change year of two adjacent

pixels was within ± 2 years, the two pixels were grouped into the same object, denoted as the *ISC objects*.

The ISC objects were then associated with the official land use maps. If an ISC object was partially or fully co-located with an undeveloped land parcel, or it was associated with two different types of developed land use, the object's actual land use was not determined as it might be mislabeled. Instead, these objects were categorized as members of the sets for predicting by the classification model as their land use type might be mislabeled on the official land use maps. There were 31,407 ISC objects in the study area: 10,485 as residential, 967 as non-residential, 2,087 as undeveloped, and 7,812 as mixed, covering 300 km^2 in total (Figure 4-3). For every type, the majority of ISC objects are all small parcels less than 0.002 km^2.

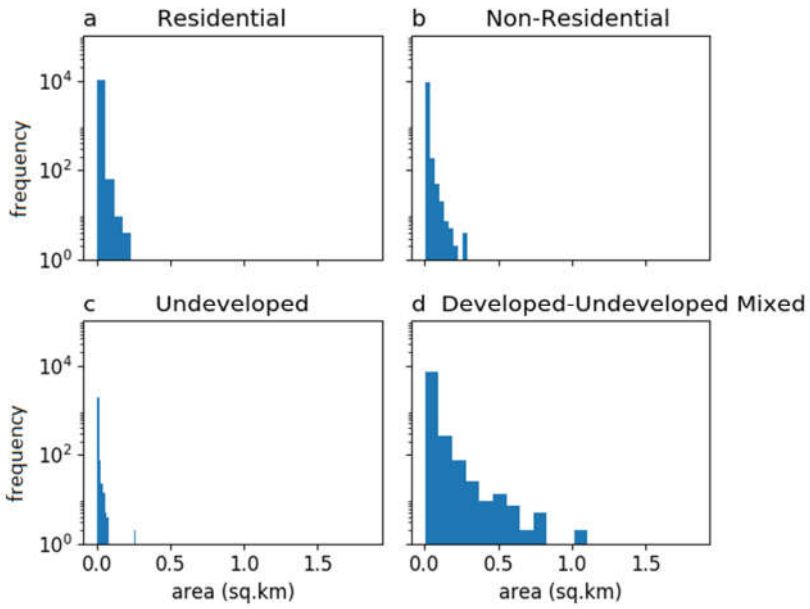


Figure 4-3 Frequency distribution of ISC objects by area

4.2. Building Physical Signatures for ISC Objects

As an ISC object is a set of pixels in each ISC-ACM layer, physical signatures can be derived from the ISC-ACM layers. Five basic statistical metrics for pixel values of an ISC object in each layer were calculated: *minimum, maximum, mean, median,* and *standard deviation*. In addition, the change magnitudes and change durations for each object were grouped by the change years, and the same five statistical metrics for these two properties in each year were calculated. In addition, three morphological metrics were also added as part of the physical signature: *perimeter, area,* and the *perimeter-area ratio* (Herold, Scepan, and Clarke 2002).

4.3. Linking Tweets to ISC Objects

Georeferenced tweets are utilized as the proxy for human activities. Before deriving temporal activity signatures, tweets from user accounts that potentially used location spoofing were removed. Location spoofing (Zhao and Sui 2017) is a technology that allows users to replace their real location by a predefined false location while using services on a mobile device, such as Twitter. It was observed that some accounts had only one or very few locations for posting a massive number of tweets. Therefore, a simple rule that removes tweets from accounts whose tweets coming from a single location takes more than 40% of their all tweets is employed to remove such spoofed tweets, excluding approximately 8% of the data set.

The remaining tweets were associated with the derived ISC objects by their location relationships. 11,633 ISC objects had co-located tweets, which accounted for 75.6% of the total area covered by all ISC objects. The ISC objects with less than seven tweets were further excluded from training, as a smaller number of tweets did not allow

for building a reliable activity signature. This filtering retained 4,694 ISC objects that accounted for 58.8% of the total area covered by all ISC objects.

4.4. Preparing Training and Validation Sets

After the above screening, there were 1,297 residential ISC objects covering 20.9 km^2, and 1,223 non-residential ISC objects covering 24.2 km^2. These objects were used as the *Training Set* for building the classification model that all ISC objects fully fall into one single developed land use parcel, i.e. residential or non-residential. The remaining 2,174 ISC objects were labeled as fully or partially undeveloped land by the official land use maps but were identified as developed by the ISC-ACM. Therefore, these ISC objects were left for prediction by the trained model, as *Application Set A*. Any ISC objects with less than seven tweets or no tweets were categorized as an independent set denoted as *Application Set B.* This Application Set B would be labeled by another classifier using the same training set, but only using the physical signature for classifying. 100 ISC objects were among the unidentified ISC objects that were randomly selected from both Application Set A and Application Set B, and denoted as *Validation Set A* and *Validation Set B* respectively. Their actual land use was manually checked on Google Maps and Google Street View as the ground truth. In the Validation Set A, there are 50 residential and 50 non-residential objects, while Validation Set B was comprised of 59 residential and 41 non-residential objects.

4.5. Building Activity Signatures for ISC Objects

Typically, the geographic units with the same land use are assumed to have similar activity signatures, thus can be used for classification. Two types of activity

patterns: temporal patterns and topic patterns were derived from the tweets for ISC objects in the Training Set and in Application Set A.

The temporal patterns of tweets in an average week are frequently used as activity signatures to characterize land use (Frias-Martinez, Soto, et al., 2012). The minimum time unit was determined to be one hour. Tweets were aggregated by day of week first, regardless of the calendar date. Three metrics were then derived: *hourly tweet volume*, *hourly user entropy*, and *hourly user volume*.

The hourly tweet volume was defined as:

$$V_{o,d,h} = \sum^{U} T_{o,u,d,h}$$

where o is the ID of an ISC object; u is the ID of a Twitter user; U is the set of user IDs; d is day of week; h ranges from 0 to 23 such that 0 represents one-hour interval between 0:00-1:00 a.m.; $T_{o,u,d,h}$ represents the total number of tweets from a unique user in an ISC object within the one-hour interval; and $V_{o,d,h}$ represents the hourly tweet volume. Generally, residential places have lower volume during week hours while non-residential places have the opposite pattern.

It has been observed however, that human behavior has a bursty nature. For example, for online behaviors, an individual may conduct some activities in a short time interval separated by a long period of waiting time, e.g. posting a large number of tweets in a short time and then waiting for a longer time before tweeting again (Barabasi, 2005; Vázquez et al., 2006). Therefore, the hourly tweet volume cannot sufficiently characterize the actual activity number in the signature as a bundle of bursty tweets may represent a single activity. Therefore, a Shannon Entropy measure (Michael Batty, 2010a; Longley & Adnan, 2016) that is commonly used for characterizing the

activity diversity is employed. Similar to hourly tweet volume, hourly user entropy thus was defined as:

$$H_{o,d,h}(U) = -\sum^{U} p(T_{o,u,d,h}) log_b^{p(T_{o,u,d,h})}$$

where $H_{o,d,h}(U)$ is the Shannon Entropy of users located at an ISC object *o* during the hourly interval h on the day of week d. $p(T_{o,u,d,h})$ is the proportion of tweets from a user among the total tweets at the same ISC object during the same hourly interval on the same day of week. It is expected that non-residential places shall have higher Shannon Entropy than residential places for users since different people may stop by and leave their digital footprint online in these places.

Hourly user volume counts the user presence at a place within an hourly interval only once and thus represents both volume and diversity. It was defined as:

$$U_{o,u,d,h} = \begin{cases} 1, & if\ T_{o,u,d,h} > 0 \\ 0, & if\ T_{o,u,d,h} = 0 \end{cases}$$

$$UV_{o,d,h} = \sum^{U} U_{o,u,d,h}$$

where $UV_{o,d,h}$ is the hourly user volume; $U_{o,u,d,h}$ represents if a user tweets in an ISC object within a specific time interval. This can reduce the effect of potential bursty tweeting activities, and can differentiate situations involving no tweets *versus* having all tweets from one single user, which cannot be characterized by the Shannon Entropy.

Single Topic LDA (ST-LDA, Hong, et al., 2016) was utilized in this study as it is particularly designed for modeling topics in tweet text and has been used for analyzing human activities (Lingzi Hong et al., 2017). The model further assumes that each tweet is associated with a single latent topic that achieves the maximum

probability to match the tweet text. 100 topics were derived from the full tweet date set. A sample of the discovered topics is displayed in Figure 4-4. Each tweet was labeled by the topic index. The counts of topics were further aggregated to each ISC object based on the spatial relationship between the ISC objects and georeferenced tweets.

Figure 4-4 Two samples of latent topics derived from the tweet set. Font sizes correspond to word weights in probability distribution. It can be interpreted that Topic a is associated with hair cutting activities and Topic b is about dinner.

4.6. Training and Classification

This study employed Random Forests algorithms (Breiman, 2001; Ho, 1995) as the main classification algorithm, since they are robust to high-dimensional feature datasets such as the feature set formed by the temporal metrics in this study, and in general may have better performance over other classifiers as reported in other work (Caruana & Niculescu-Mizil, 2006). The main idea of Random Forests classifiers is to build a collection of decision trees with each being trained independently based on a randomly selected subset of original features. At the classification stage, each decision tree predicts the class label of a record independently, with the final label of the record being determined by majority voting among the prediction label set. Implementation of Random Forests in *scikit-learn* (Pedregosa et al., 2011) was used.

For building the Random Forests model, 10-fold cross-validation was used to evaluate the performance of classification model building on the training dataset. Ten is considered an optimum number for cross-validation for comparing model performance due to relatively low inter-fold bias and variance (Kohavi, 1995). 10-fold cross-validation splits the training set into 10 equal-size folds and uses nine folds to build a classification model and one remaining fold to evaluate the model. For performance evaluation, such as accuracy, ISC objects, rather than the areas, are used as the basic unit. In this way, the best set of parameters for a Random Forests model on this data set can be found.

To evaluate the models in the cross-validation process, accuracy, Cohen's Kappa coefficient (Cohen, 1960), precision, recall, F1-score, and the area under the receiver operating characteristics curve (AUC) are used to evaluate a classifier comprehensively. Precision is the percentage of real positive records in the dataset that are predicted as positive by the classifier. Recall is the percentage of records that are correctly predicted as positive in all positive records. F1-score is the harmonic average of precision and recall (Han et al., 2012). AUC shows the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance (Fawcett, 2006).

Two training processes were conducted on the same training set with different signature combinations: both physical signatures and activity signatures were used for the first classification model to identify Application Set A that included 1403 attributes (Table 4-1); only physical signatures were used for the second classification model to identify the Application Set B. In general, more trees in a Random Forests result in

higher classification accuracy, but 64-512 trees are sufficient to achieve a good performance (Oshiro, Perez, & Baranauskas, 2012). In this analysis, each Random Forests model was composed of 256 trees. The classification performances of the two models were evaluated by the 10-fold cross-validation. After the two application sets were classified, their corresponding testing sets were also applied to evaluate the two models respectively as the proxy of all objects.

Table 4-1 Feature groups and their index for the classifier model (ln stands for natural logarithm).

Signature type	Feature group	Index
Activity signature	Hourly tweet volume	0~167
	ln(Hourly tweet volume)	168~335
	Hourly user entropy	336~503
	ln(Hourly user entropy)	504~671
	Hourly user volume	672~839
	ln(Hourly user volume)	840~1007
Physical signature	Perimeter, area, perimeter-area-ratio	1008~1010
Activity signature	Topic counts	1011~1110
Physical signature	Statistics of change year	1111~1116
	Statistics of change magnitude	1116~1121
	Statistics of change duration	1121~1126
	Statistics of change magnitude per year	1126~1264
	Statistics of change duration per year	1265~1402

5. Results

5.1. Model Performances of 10-fold Cross-validation

For the 10-fold cross-validation of the classifier using both physical and activity signatures, their average accuracy was 0.81 with a standard deviation 0.03, and the best accuracy is 0.87. The average Kappa was 0.62 with a standard deviation 0.06, which falls in the range of substantial agreement (Landis & Koch, 1977). The average AUC is also 0.81 with a standard deviation 0.03. In addition, the precision and recall values

were balanced (Table 4-2), meaning that the high accuracy was not achieved by consistently predicting all objects as one single type.

Table 4-2 Detailed classification report of a selected cross-validation on features from both physical and activity signatures (accuracy: 0.87, Kappa coefficient: 0.74, AUC: 0.87)

	Precision	Recall	F1-score
Non-residential	0.89	0.84	0.86
Residential	0.85	0.90	0.88
Average	0.87	0.87	0.87

According to two additional 10-fold cross-validations on the classifiers using the same parameters, using the two signatures separately achieved slightly worse performance (Table 4-3). The performance metrics of using both signature combinations are all significantly higher than the results based on testing them independently using a t-test (p-value < 0.01). This suggests that the activity signature does contribute extra information into the land use classification modeling.

Table 4-3 Model performance of 10-fold cross-validation on three signature combinations.

Signature combination	Average Accuracy	Average Kappa	Average AUC
Physical + activity	0.81	0.62	0.81
Physical only	0.77	0.54	0.77
Activity only	0.75	0.49	0.75

A feature importance analysis was also conducted on the combination of using both types of signatures. This analysis suggests whether a feature is informative for the classification task (Breiman, 2001). There were features from both types of signatures contributing relatively more to the classification result than the rest (Figure 4-5). The mean feature importance associated with physical signatures was higher than the mean feature importance of activity signatures ($p < 0.01$), indicating that the features in the physical signature groups were more informative to differentiate residential and non-residential land use. Among the three metrics in the activity signature, hourly tweet volume and hourly user entropy were found to be more informative than hourly user

volume (both with $p < 0.01$). Highly-ranked topic features were those that refer to common activities with strong spatial contexts associating with residences, such as topics about sleeping and gaming.

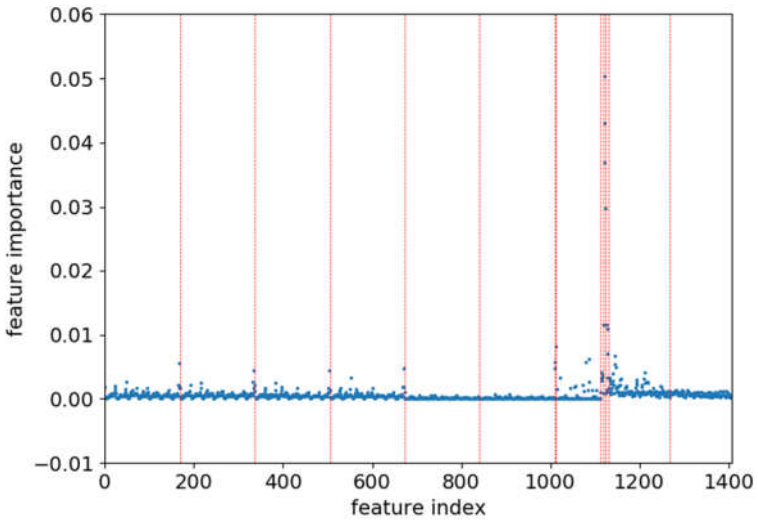


Figure 4-5 Relative feature importance of the physical signature and activity signature. The feature groups and indexes are the same as Table 4-1.

5.2. Model Performance on Validation Sets

By evaluating the 100 randomly selected ISC objects in Validation Set A predicted by the Random Forest model using the full training set and the same model parameters, the overall accuracy was 0.87, with a Kappa coefficient 0.74 and an AUC 0.87. The three overall performance metrics were slightly better than most results in 10-fold cross-validation while the validated accuracy was still in the range of two standard deviations of the mean 10-fold cross-validation accuracy. However, the model had a slightly lower performance regarding the precision of the non-residential type and the recall of the residential type than the results from the 10-fold cross-validation, even though the number of residential ISC objects was larger than the number of non-residential ISC objects in the training set.

Table 4-4 Detailed classification report on the Validation Set A based on the 100 validation

ISC objects

Accuracy	0.87		
Kappa	0.74		
AUC	0.87		
	Precision	**Recall**	**F1-score**
Non-residential	0.81	0.96	0.88
Residential	0.95	0.78	0.86
Avg.	0.88	0.87	0.87

The accuracy measurement of area is subject to the areal extent of each object. Following the recommended practice for area-adjusted accuracy estimation in the remote sensing field (Olofsson et al., 2014), the estimated accuracy and estimated error matrix (Olofsson, et al., 2013) were calculated in order to demonstrate the difference (Table 4-5). The decreased overall accuracy may subject to the misclassification of ISC objects with large areas.

Table 4-5 Area-adjusted accuracy and error matrix on the 100 validation ISC objects in Validation Set A. The margin of error is based on 1.96 times of standard error of the estimators, which provides 95% confidence.

Estimated overall accuracy	0.81 ± 0.01	
	Estimated precision	**Estimated recall**
Non-residential	0.75±0.01	0.98±0.003
Residential	0.96±0.01	0.61±0.006

For Validation Set B, the overall accuracy was 0.54, with a Kappa coefficient 0.03 and an AUC 0.51 (Table 4-6). The area-adjusted performance estimators were better than the object-based estimators (Table 4-7). This was likely due to the large number of small objects in Application Set B and Validation Set B (objects with less than two pixels were 60% of the count, but accounted for 17% of the overall area in Validation Set B). Therefore, the area-adjusted accuracy was a little better, but still much lower than the result of the model utilizing both physical and activity signatures.

Table 4-6 Detailed classification report on the Validation Set B based on the 100 validation ISC objects

Accuracy	0.54
Kappa	0.02

AUC		0.51	
	Precision	**Recall**	**F1-score**
Non-residential	0.64	0.62	0.63
Residential	0.38	0.41	0.39
Avg.	0.55	0.54	0.54

Table 4-7 Area-adjusted accuracy and error matrix on the 100 validation ISC objects in Validation Set B. The margin of error is based on 1.96 times of standard error of the estimators, which provides 95% confidence.

Estimated overall accuracy	0.72 ± 0.04	
	Estimated precision	**Estimated recall**
Non-residential	0.80±0.04	0.78±0.001
Residential	0.55±0.09	0.58±0.03

6. *Sprawl of Residential vs Non-Residential Land in the DC-Baltimore Metropolitan Area*

The sprawl by built-up areas in the DC-Baltimore metropolitan area over time as computed using our approach was mapped (Figure 4-6). Generally, new developed non-residential places cluster along main transportation corridors, while residential neighborhoods scatter around these non-residential places. In terms of the total area, the overall increase of residential areas was slightly smaller than for non-residential areas in the 1986-2008 period (Table 4-8). Using the indicated changed year in ISC-ACM Change Year, the temporal characteristics of total land use sprawl was profiled in Figure 4-7. The overall time in which sprawling of residential and non-residential changes occurred followed the same trend as observed. The increase of non-residential areas was greater than that of residential after 1996.

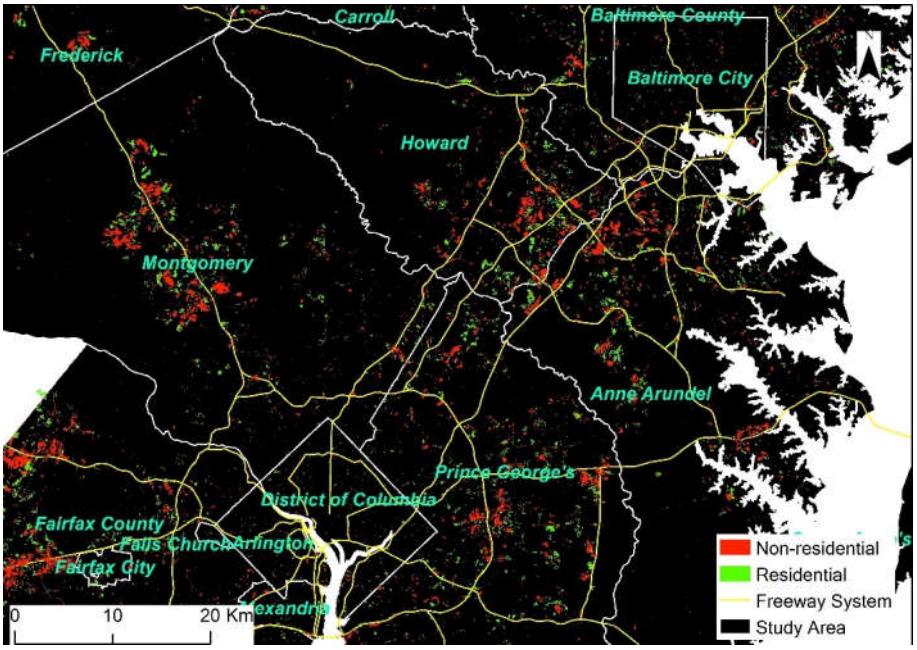


Figure 4-6 Non-residential and residential area developed between 1986 and 2008 in Washington D.C.-Baltimore region by the three sub data sets. The values of the Training Set are the ground truth from land use maps. The values of the other two labeling sets are based on modeling prediction.

Table 4-8 Areas of residential and non-residential using the same approach as Figure 4-6. The unit of the values is km^2. The margin of error is based on 1.96 times of standard error of the estimators, which provides 95% confidence.

	Residential	Non-Residential	Total
Training: Truth	20.63	24.14	44.77
Application Set A: Predicted	40.58±1.19	91.07±1.19	131.65
Application Set B: Predicted	64.16±9.37	74.20±9.37	138.36
Total	125.37±10.56	189.41±10.56	314.78

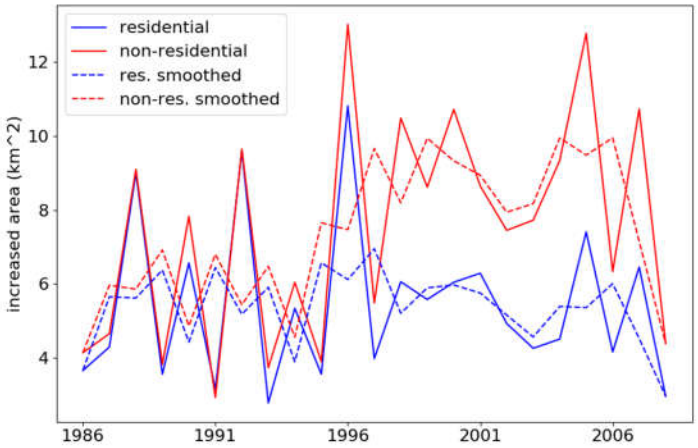


Figure 4-7 Residential and non-residential area increases by year using the same approach as Figure 4-6. The smoothed curves are based on the average of a three-year moving window.

The yearly increases of residential and non-residential areas in each administrative entity showed that sprawl mainly occurred in seven counties in Maryland including, Anne Arundel, Baltimore, Frederick, Harford, Howard, Montgomery, and Prince Georges' County, and Fairfax County, VA (Figure 4-8). It was also observed that the increase in non-residential areas surpassed the increase in residential areas after 1996 for the eight counties, except Fairfax County, where this increase started earlier, in 1988. For Montgomery County, this extra increase can be explained by the I-270 Technology Corridor stretching from Bethesda, MD to Rockville, MD, where over 18,000 business establishments have located, offering 72% of Montgomery County's total employment, while 30% of the employees lived outside of the County, and most housing growth was estimated to be multi-family as of 2007 (Tate, et al., 2007). For Fairfax County, the amount of increase could be explained by similar reasons, as there is the Dulles Technology Corridor connecting cities in Fairfax County, VA and involving communities such as Tysons Corner, Reston, Herndon, Sterling and Ashburn.

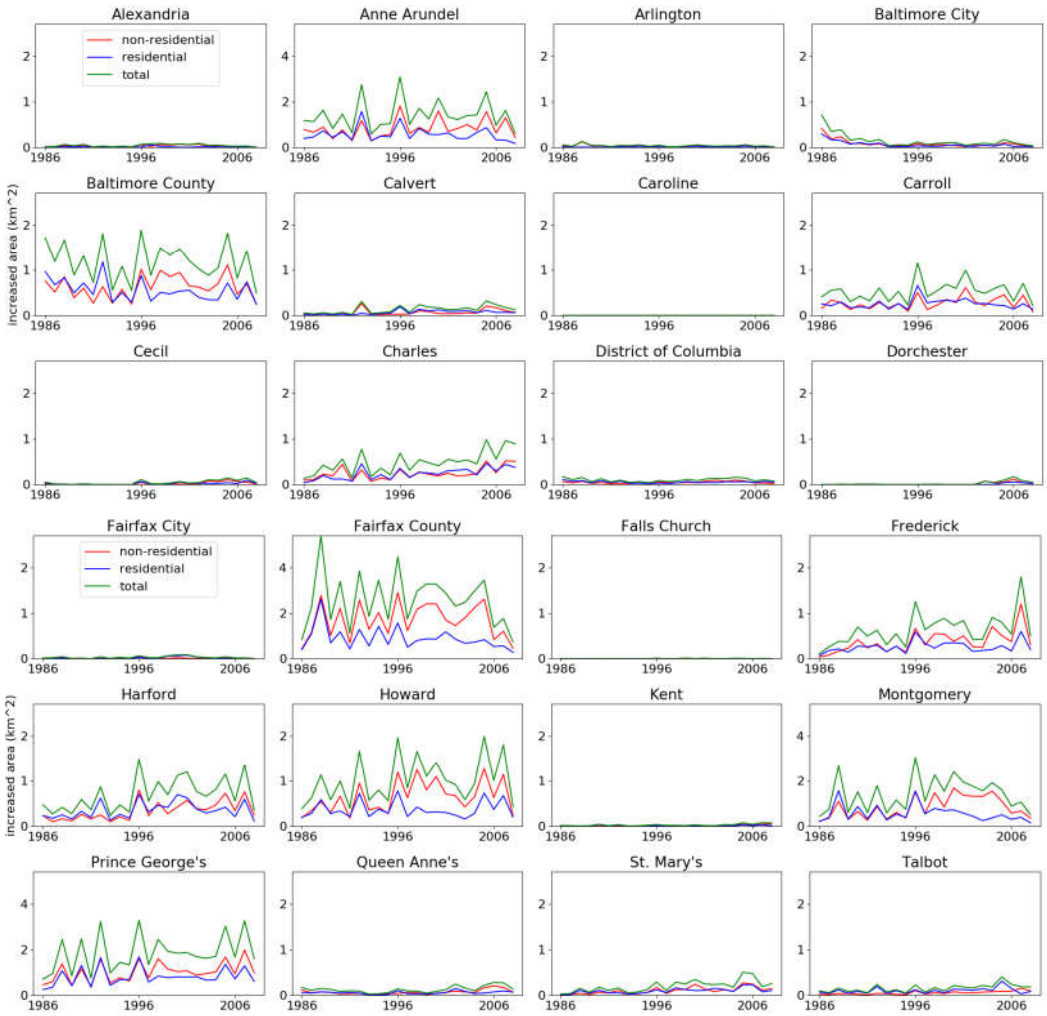


Figure 4-8 The increased areas of non-residential and residential places by administrative entities and years by using the same approach as Figure 4-6.

7. Discussion

In this study, we propose a framework that integrates both remotely sensed imagery and socially sensed human activities data to identify detailed urban land use. The output of the framework not only maps land use details spatially, but also profiles the trajectories of different land use types over time, which can contribute a better understanding of the evolution of urban development as a complex system. The framework minimizes the dependence on ground surveying GIS data sets such as street

network maps and land parcel footprint maps that are costly in terms of both time and finance. Since the original data sources, the Landsat imagery and georeferenced tweets, are free to access, the framework has the potential to be applied to larger areas, especially in developing countries, where cities are undergoing fast urbanization and land use mapping may not be able to keep up. For municipalities or counties in the US with zoning or land use maps, the output of this framework may help to address mapping errors in current County maps, such as the case in Prince George's County, MD where a primary school founded in 2007 is still marked as pasture in the official land use map (Figure 4-9).

Figure 4-9 The Friends Community School on the 2010 Maryland Land Use Land Cover Map (left) and on the Google Maps (right). The land parcel that the school locates (marked as the red star in the official land use map) is mislabeled as pasture, although it was converted to school in 2007.

This framework utilizes remote sensing imagery to model the physical signature of land cover and georeferenced tweets to model activity signatures associated with different land use types. The comparison of classification models shows that the area-adjusted accuracy of the model when using both signatures is about 0.10 higher than if the model uses a physical signature alone. This improvement is based on the distinguishable residential vs. non-residential landscape pattern in the Washington

D.C.-Baltimore region. This region has been experiencing suburbanization at a high rate where single-house communities with cul-de-sac designs are significantly different from commercial parcels in terms of morphology, and the magnitude of impervious surface. This is not necessary true for cities in other regions with compact urban land parcel patterns, e.g., New York City, Beijing, China, and Manila, Philippines. Activity signatures can bring more value there to differentiate the land use of parcels. In addition, different types of non-residential often have similar high impervious surface cover, that may be more difficult to distinguish using the physical signatures alone.

The topic features extracted as part of the activity signature analyses are observed to have high importance in the classification model. This implies that topic features could be further investigated for advanced classification tasks, for example, classifying detailed non-residential land use types. Conceptually, each detailed non-residential land use type has unique corresponding activity types. For example, the main activities associated with schools are teaching and learning, which are different from the main activities associated with shopping malls or grocery stores. If such activity information can be retrieved, it is possible to use such information for identifying more detailed non-residential land use types.

The tradeoff of this study is to define the spatial footprint of places from segmenting only remotely sensed imagery. Currently, image segmentation-based place footprints do not perfectly match with ground truth. As an example, using Application set A, parcel A, a commercial complex, and parcel C, a residential community with cul-de-sacs, are well identified (Figure 4-10). However, parcel B, which is based on a combination of commercial buildings in the north and some residential buildings in the

south, was labeled as non-residential. If a pixel in the commercial buildings part had been selected as a testing pixel, the label would be correct, otherwise though, the attribution would not be correct. Unlike land cover objects, it is difficult to define the ground truth of a land use object, as a land use object may involve several land cover types. The boundary of a land use object may also be subject to a person's feelings about a place (Tuan, 1979).

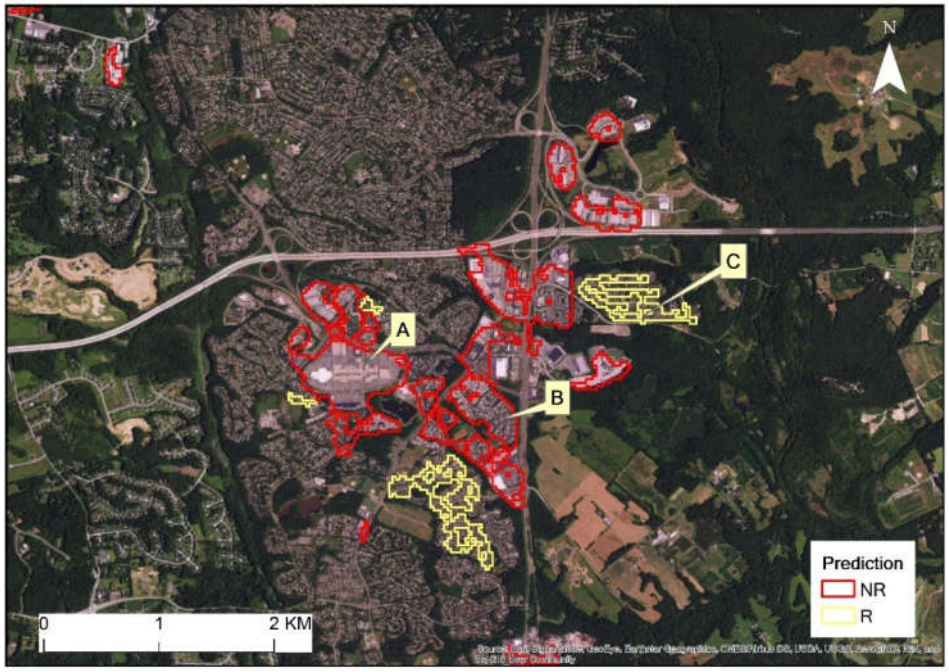


Figure 4-10 Detailed ISC object classification result of Application Set A near Bowie, MD. R: residential. NR: non-residential.

A more challenging issue, however, is the decline in numbers of GPS-tagged tweets in the georeferenced tweets. The proportion of tweets having exact GPS coordinates dropped dramatically after April 2015 in the collected data set (Figure 4-11). The remaining tweets are tagged by a nearby place but have no GPS coordinates to show the exact location. Due to this issue, the activity signatures can be difficult to derive, or need a much longer time period before collecting sufficient data. Other solutions can involve using other types of socially sensed data that can represent human

activities, such as vehicle trajectories, check-in records, and CDRs, though these data sets may require licensing.

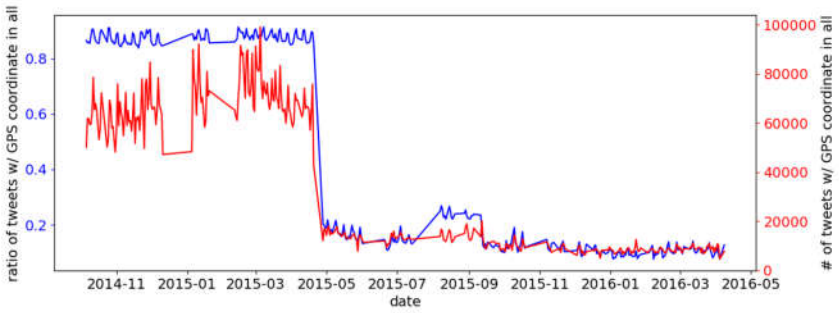


Figure 4-11 The ratio of tweets with GPS coordinates in the all tweets collected via Twitter Public Streaming API

8. *Conclusion*

An innovative framework has been developed to identify land use types for places in metropolitan areas based on modeling the physical and activity signatures of places using remotely sensed imagery and socially sensed human activity data. The framework can provide a land use map with over 80% accuracy. We also showed that introducing the activity signatures of places could improve the classification performance, compared to using features derived solely from remotely sensed imagery. Using the land use map produced by the framework, we observed that 125.37±10.56 km2 new residential land use and 189.41±10.56 km2 non-residential land use occurred in the Washington D.C.-Baltimore metropolitan area during the 1986-2008 period. The analysis of the temporal profile of urban sprawl showed that the increase in non-residential land use surpassed residential land use during the same period for this area. The analysis on the temporal profile of urban sprawl showed that non-residential land use surpassed residential land use during the same period in the same area. For future work, the assumptions on data gaps and modeling could be tested to more fully

understand the impacts of the activity signatures as well as data quality, e.g., the uncertainty of GPS coordinates retrieved by smartphones, for identifying land use types. We will also be focused on evaluating the potential applicability of the proposed framework for classifying different non-residential land use types. In doing so, the complex nature of urban development would be better understood.

Chapter 5: : Conclusions and Future Work

1. Conclusions

Increasingly, urban studies model cities as complex systems composed of different dynamic processes involving interactions among people, infrastructure, information, capital, etc. Human activities are a critical part of a city, whose patterns can be employed as a lens to gain insights into the complexities of urban dynamics. This dissertation contributes a set of methods for monitoring the impact of urban land-use structure on geographic patterns of human activity patterns as an important part of urban dynamics. Socially sensed data, e.g., georeferenced tweet data derived from Twitter as used in this dissertation, are known as potentially biased sources in terms of representing the demography of population. They are also biased for representing urban residents. However, by utilizing such a freely-accessible Big Data set, its large volume and good spatiotemporal coverage and detail make it a good proxy for human activities that are ongoing in a city. This dissertation presents three studies that utilize different perspectives to analyze the city as a dynamic and complex system, with an ultimate goal of creating pathways that can contribute empirical results to build knowledge about the science of cities.

The first of three studies in this dissertation models the associations between socioeconomics and mobility using the Washington, D.C. metropolitan area as a case study, and applies the learned associations for inferring geographical patterns of socioeconomic status (SES) through the sole use of human activity data. The second study designs and implements a semi-automated workflow to sense details of urban activities using socially sensed Twitter data. The third study reveals the relationship

between different land-use types and the spatiotemporal variation of activities, and tests if activity information collected through socially sensing platforms can be used as a source for mapping urban land use with remote sensing imagery, providing a new way forward for automated regional mapping tasks in the future.

Even as we are aware of some of the limitations of the empirical data used to represent individuals' activity patterns, the first study confirms the extremely heterogeneous spatial pattern of human activities in a city. This study further demonstrates that for the metropolitan area used as a case study (Washington, D.C.), there is no simple universal correlation between SES and mobility and that a local population with high SES does not guarantee correspondingly large mobility, while populations with lower SES also demonstrate a large activity space especially where public transportation options are available. Urban geography also appears to influence residents' lifestyles in that residents of suburban areas have higher spatial dispersion but lower diversity than residents in the downtown urban core. The first study applies network analysis to analyze spatial interactions between different places to infer the landscape of the population's SES returning a moderate level of agreement with the actual SES pattern. This approach shows promise as an alternative for estimating neighborhoods with different SES in cities where census data are not available.

The second study undertaken for this dissertation applies natural language processing (NLP) technology on activity topics extracted from the content of georeferenced tweets from three U.S. cities to identify different types of activities in cities. The derived topics are semantically, temporally, and spatially related to the activities. The derived activity topics are used to profile the unique social functions of

a neighborhood. The modeling results and statistical analyses show that similarities between neighborhoods based on the activity profiles are consistent between inter-city and intra-city with respective to the similarity from a social, economic and historical perspective. This investigation successfully characterizes neighborhoods from three U.S. cities based on the signatures of activities tweeted by individuals.

The last study presents a framework to infer land use types in an urban context through the integration of data acquired by remote sensing and social sensing. The framework uses remotely-sensed satellite imagery to model the physical signatures of land use. A georeferenced Twitter data set is employed to model the activity signatures of places drawing on the knowledge about the heterogeneous patterns of human activities in space and time from the first two studies. The framework is applied to map land use and its change in the Washington, D.C.-Baltimore metropolitan area between 1986 and 2008, and achieves over 80% accuracy for differentiating residential and non-residential land uses. This analysis also shows that the development of residential and non-residential use types has not been consistent during the studied period, and that non-residential land use surpassed residential use in the studied area after 1996.

However, it is still an open question how the bias in representing sociodemographic groups may influence the generality of the findings in this dissertation, particularly in quantitative ways due to the lack of complete knowledge about activities. A potential solution is to conduct a survey on a sample of Twitter users to calibrate the weights of different sociodemographic groups for representing the observed activities from the georeferenced tweets. Aligning other socially sensed data sources that may have a larger user group, e.g., call detail records from mobile phone

carriers with the same methodology could be another way to evaluate the influence of bias quantitatively.

Data and algorithm ethics in relation to the privacy of individuals can be another concern that applies to the studies in this dissertation as well as to other socially sensed data sources and areas. Individuals can be easily differentiated from each other with only a few spatiotemporal data points (de Montjoye, et al., 2013) or their profiles (Quercia, et al., 2011). On the other hand, users are often aware of privacy issues and take action through their preferences for sharing place-based information and the exposure of their privacy while using social media with location-based services (Benisch, et al., 2011; Lin et al., 2012; Zhou & Li, 2014). The importance of privacy, has also been discussed as part of a more general concern for balancing personal privacy with the benefit of promoting social studies research and gaining more spatiotemporal details (Elwood & Leszczynski, 2011). In this dissertation, the concern for privacy is supported through anonymizing Twitter users and only studying collective behavior patterns. For future research, data ethnics should be acknowledged and acted upon, and studies targeting individual users should be avoided.

2. Significant Results

Resulting from the research undertaken for this dissertation, there are a number of innovative and significant findings as well as innovations in the methodologies applied.

Innovation 1: For the first study, quantitative metrics including radius of gyration and entropy are employed to measure and reveal the spatial dispersion and diversity of human mobility and the association between the metrics and the

populations' SES derived from the traditional social area approach. The results of this study (presented in Chapter 2) show the complex relationships between SES and mobility where individuals with high SES in Washington, D.C. do not necessarily exhibit higher mobility than other groups, a result that is new compared to previous studies relying on small-sample surveys in other cities.

Innovation 2: In the second study (Chapter 3), topic modeling, particularly a variant of latent Dirichlet allocation (LDA), is applied as the core step of a novel semi-automated knowledge discovery pipeline to derive and extract activity topics from georeferenced tweets. The derived topics are validated from semantic, temporal, and spatial perspectives, showing how this new pipeline can provide more details of spatiotemporal patterns of different activity types in a city with free-accessible data than previous studies have revealed.

Innovation 3: The second study also provides an innovative approach to characterize neighborhoods by human activity signatures and measure the similarity of neighborhoods using the activity signatures. This provides a new perspective as compared to conventional approaches such as sociodemographic signatures that are mainly based on residents' socioeconomic status from census. The new activity-based approach can capture similarity based on social, economic, and historical dimensions that cannot be characterized by sociodemographic signatures.

Innovation 4: In the third study (Chapter 4), an automated workflow that includes an innovative integration of remote sensing imagery with a socially sensed data set is implemented for mapping detailed urban land use over time. The remote sensing product is used for modeling the features of land cover, while socially sensed

data are used for modeling the features of human activities. The improvement that arises from combining these two types of features is observed and compared to the conventional remote sensing approach with the same machine learning classifier for classifying land use in the Washington, D.C.-Baltimore metropolitan.

Innovation 5: Using the mapping results of the innovative mapping workflow described in Chapter 4, the process of land use change in the Washington, D.C.-Baltimore metropolitan between 1986-2008 is investigated. The results of the analysis are that the total new residential land use is approximately 125.37±10.56 km^2, while the new non-residential land use is 189.41±10.56 km^2. In addition, this research has revealed that non-residential land use surpassed residential land use beginning in 1996 in terms of area. These results may be related to the boost felt by local businesses situated in the two growing technology corridors (i.e., I-270 corridor and Dulles Airport corridor) in the capital area.

3. *Future Work*

Social sensing is still a new science in Geography. It enables sensing *and* studying human activities over a large area with the opportunity to expose fine-grained temporal and spatial details. In this dissertation, three studies have been conducted to understand and further the science of human activity and urban geography. However, as a new science, and due to some of the data limitations and the scope of these studies, there remains a number of open research topics to address in the future.

As discussed in the chapters of this dissertation, a major issue with socially sensed data is the potential bias with respect to representing the population's activity patterns. One possible future direction is to collect socially sensed data from more than

one source, such as Twitter, call detailed records, taxi trajectories, etc. These data sets are sampled with different biases in terms of demography. Investigating how these data can be fused for modeling human activity, or how the same methodology could be applied to these different data sets, then new results and insights might be gained.

For research on modeling associations between mobility and SES, future research could investigate how to design finer-grained models that model individuals' socioeconomic status independently while also capturing interactions between, for example, census enumeration areas. For this dissertation research, individuals are treated as being uniform, however, observed movements could be calibrated by the diverse demographies of the source enumeration units. The influence of physical geography on human mobility could also be considered as a factor in the moderate agreement found between the estimated SES of this research and the SES patterns using other social area approaches. Calibrating the influence of physical geography and physical features may improvement the agreement and could be a topic for future study.

Relating to the work presented in Chapter 3, future studies could investigate further the ontology of the derived activities in order to determine what taxonomy is the best for describing the variety of activities in a city. This would also contribute to understanding how people develop a sense of place through the activities in and around places. The methodology presented here can be applied to additional cities to capture the activity signatures in different cities and implement a formal recommendation system for suggesting similar neighborhoods from a set of different cities. So far, the similarity matrix among neighborhoods in the three cities investigated in Chapter 3 is already to be used for neighborhood recommending. This type of system, as an App for

the public or a similarity index on lifestyle for real estate industry, will be helpful for people who want to maintain their current lifestyle and activities when moving to a new city.

Future work could also consider additional data sources, e.g., products of a Light Detection and Ranging (LiDAR) systems or other trajectories datasets that can be integrated into a framework to provide more physical features and activity features of urban places that in turn would improve the classification accuracy. Another possibility is to apply such a framework to differentiate more detailed land uses, for example, differentiating commercial and public land uses within the category of non-residential land use. Future work involving more detailed validation for the processing of land use change is also needed to understand the drivers behind the land use changes. This framework has the potential to be utilized by planning departments, especially in areas and countries without mature land use monitoring systems or strong zoning systems to map and trace the land use changes.

Bibliography

Adams, B., McKenzie, G., & Gahegan, M. (2015). Frankenplace: interactive thematic mapping for ad hoc exploratory search. In *Proceedings of the 24th International Conference on World Wide Web - WWW '15* (pp. 12–22). New York, New York, USA: ACM Press. https://doi.org/10.1145/2736277.2741137

Alsnih, R., & Hensher, D. A. (2003). The mobility and accessibility expectations of seniors in an aging population. *Transportation Research Part A: Policy and Practice*, *37*(10), 903–916. https://doi.org/10.1016/S0965-8564(03)00073-9

Anselin, L. (1995). Local Indicators of Spatial Association - Lisa. *Geographical Analysis*, *27*(2), 93–115.

Arlington County. (2015). Zoning Map, Arlington County, VA. Retrieved June 20, 2010, from http://gis.arlingtonva.us/gallery/map.html?webmap=1e4706ab574a462a8dcc6a6c182b0004

Bajardi, P., Delfino, M., Panisson, A., Petri, G., & Tizzoni, M. (2015). Unveiling patterns of international communities in a global city using mobile phone data. *EPJ Data Science*, *4*(1), 3. https://doi.org/10.1140/epjds/s13688-015-0041-5

Barabasi, A.-L. (2005). The origin of bursts and heavy tails in human dynamics. *Nature*, *435*(7039), 207–211. https://doi.org/10.1038/nature03526.1.

Barthélemy, M. (2011). Spatial networks. *Physics Reports*, *499*(1–3), 1–101. https://doi.org/10.1016/j.physrep.2010.11.002

Batty, M. (2010a). Space, Scale, and Scaling in Entropy Maximizing. *Geographical Analysis*, *42*(4), 395–421. https://doi.org/10.1111/j.1538-4632.2010.00800.x

Batty, M. (2010b). The pulse of the city. *Environment and Planning B: Planning and Design*, *37*(4), 575–577. https://doi.org/10.1068/b3704ed

Batty, M., Besussi, E., & Chin, N. (2003). Traffic, Urban Growth and Suburban Sprawl. *Centre for Advanced Spatial Analysis*, *44*(0), 0–18. https://doi.org/10.1103/PhysRevE.78.016110

Benisch, M., Kelley, P. G., Sadeh, N., & Cranor, L. F. (2011). Capturing location-privacy preferences: quantifying accuracy and user-burden tradeoffs. *Personal and Ubiquitous Computing*, *15*(7), 679–694. https://doi.org/10.1007/s00779-010-0346-0

Blaschke, T. (2010). Object based image analysis for remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing*, *65*(1), 2–16. https://doi.org/10.1016/j.isprsjprs.2009.06.004

Blei, D. M., & McAuliffe, J. D. (2008). Supervised Topic Models. *Advances in Neural Information Processing Systems 20*, 121–128. Machine Learning. Retrieved from http://arxiv.org/abs/1003.0783

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation.pdf. *Journal of Machine Learning Research*, *3*, 993–1022. https://doi.org/10.1.1.110.4050

Blessett, B. (2015). African Americans and gentrification in Washington, DC: race, class and social justice in the nation's capital. *Ethnic and Racial Studies*, *38*(13), 2402–2404. https://doi.org/10.1080/01419870.2014.987794

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), 1–12. https://doi.org/10.1088/1742-5468/2008/10/P10008

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5–32.

Brown, D. . G., Pijanowski, B. . C., & Duh, J. . D. (2000). Modeling the relationships between land use and land cover on private lands in the Upper Midwest, USA. *Journal of Environmental Management*, *59*(4), 247–263. https://doi.org/10.1006/jema.2000.0369

Brown, D. G., Carolina, N., & Hill, C. (2012). Walking Within a City. *Am J Prev Med*, *40*(3), 353–361. https://doi.org/10.1016/j.amepre.2010.11.017.A

Burby, R. J., Deyle, R. E., Godschalk, D. R., & Olshansky, R. B. (2000). Creating Hazard Resilient Communities through Land-Use Planning. *Natural Hazards Review*, *1*(2), 99–106. https://doi.org/10.1061/(ASCE)1527-6988(2000)1:2(99)

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. *Proceedings of the 23rd International Conference on Machine Learning*, *C*(1), 161–168. https://doi.org/10.1145/1143844.1143865

Cervero, R., & Kockelman, K. (1997). Travel demand and the 3Ds: Density, diversity, and design. *Transportation Research Part D: Transport and Environment*, *2*(3), 199–219. https://doi.org/10.1016/S1361-9209(97)00009-6

Cervone, G., Sava, E., Huang, Q., Schnebele, E., Harrison, J., & Waters, N. (2016). Using Twitter for tasking remote-sensing data collection and damage assessment: 2013 Boulder flood case study. *International Journal of Remote Sensing*, *37*(1), 100–124. https://doi.org/10.1080/01431161.2015.1117684

Chae, J., Thom, D., Bosch, H., Jang, Y., Maciejewski, R., Ebert, D. S., & Ertl, T. (2012). Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition. *IEEE Conference on Visual*

Analytics Science and Technology 2012, VAST 2012 - Proceedings, (July), 143–152. https://doi.org/10.1109/VAST.2012.6400557

Chalasani, V. S., Denstadli, J. M., Axhausen, K. W., & Engebretsen, Ø. (2005). Precision of Geocoded Locations and Network Distance Estimates. *Journal of Transportation and Statistics*, *8*(2), 1–16.

Chang, J., Gerrish, S., Wang, C., Boyd-graber, J. L., & Blei, D. M. (2009). Reading Tea Leaves : How Humans Interpret Topic Models. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in Neural Information Processing Systems 22* (pp. 288–296). Curran Associates, Inc. https://doi.org/10.1.1.100.1089

Cheng, Z., Caverlee, J., Lee, K., & Sui, D. Z. (2011). Exploring Millions of Footprints in Location Sharing Services. In *ICWSM* (pp. 81–88).

Christaller, W. (1933). *Die zentralen Orte in Suddeutschland (Central Places in Southern Germany)*. Jena: Gustav Fischer.

City of Alexandria. (2015). City of Alexndria 2015 Zoning Map. Retrieved June 20, 2010, from https://www.alexandriava.gov/uploadedFiles/gis/info/Zoning2015.pdf

City of Falls Church. (2015). Official Zoning District Map. Retrieved June 20, 2010, from http://www.fallschurchva.gov/DocumentCenter/View/690

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46.

Cranshaw, J., Schwartz, R., Hong, J. I., & Sadeh, N. (2012). The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City Justin. In *6th*

ICWSM (pp. 58–66).

Cresswell, T. (2014). *Place: An Introduction*. Wiley. Retrieved from https://books.google.com/books?id=6OJvBAAAQBAJ

Crooks, A., Pfoser, D., Jenkins, A., Croitoru, A., Stefanidis, A., Smith, D., … Lamprianidis, G. (2015). Crowdsourcing urban form and function. *International Journal of Geographical Information Science*, *29*(January), 37–41. https://doi.org/10.1080/13658816.2014.977905

Dalziel, B. D., Pourbohloul, B., & Ellner, S. P. (2013). Human mobility patterns predict divergent epidemic dynamics among cities. *Proceedings of the Royal Society B: Biological Sciences*, *280*(1766), 20130763–20130763. https://doi.org/10.1098/rspb.2013.0763

DC Office of Planning. (2006). Existing Land Use Maps. Retrieved October 1, 2016, from https://planning.dc.gov/page/existing-land-use-maps

De Montis, A., Caschili, S., & Chessa, A. (2013). Commuter networks and community detection: A method for planning sub regional areas. *The European Physical Journal Special Topics*, *215*(1), 75–91. https://doi.org/10.1140/epjst/e2013-01716-4

de Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the Crowd: The privacy bounds of human mobility. *Scientific Reports*, *3*. https://doi.org/10.1038/srep01376

Dovey, K. (2012). Informal urbanism and complex adaptive assemblage. *International Development Planning Review*, *34*(4), 349–368. https://doi.org/10.3828/idpr.2012.23

Duggan, M. (2015). *Mobile messaging and social media 2015*. *Pew Research Center*. Retrieved from http://www.pewinternet.org/2015/08/19/mobile-messaging-and-social-media-2015/

Eisenstein, J., & O'Connor, B. (2010). A latent variable model for geographic lexical variation. *Proceedings of the 2010 ...*, 113–120. https://doi.org/10.1.1.173.3302

Elwood, S., & Leszczynski, A. (2011). Privacy, reconsidered: New representations, data practices, and the geoweb. *Geoforum*, *42*(1), 6–15. https://doi.org/10.1016/j.geoforum.2010.08.003

Ester, M., Kriegel, H., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *KDD'96 Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (pp. 226–231). Elsevier. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/B9780444527011000673

Fairfax County GIS & Mapping Service Branch. (2015). Zoning, Farifax County, VA. Retrieved October 1, 2016, from http://data-fairfaxcountygis.opendata.arcgis.com/datasets/zoning

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Fonte, C. C., Bastin, L., See, L., Foody, G., & Lupia, F. (2015). Usability of VGI for validation of land cover maps. *International Journal of Geographical Information Science*, *29*(7), 1269–1291. https://doi.org/10.1080/13658816.2015.1018266

Frias-Martinez, V., Soguero, C., & Frias-Martinez, E. (2012). Estimation of urban commuting patterns using cellphone network data. In *Proceedings of the ACM*

SIGKDD International Workshop on Urban Computing - UrbComp '12 (p. 9). New York, New York, USA: ACM Press. https://doi.org/10.1145/2346496.2346499

Frias-Martinez, V., Soto, V., Hohwald, H., & Frias-Martinez, E. (2012). Characterizing Urban Landscapes Using Geolocated Tweets. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing* (pp. 239–248). IEEE. https://doi.org/10.1109/SocialCom-PASSAT.2012.19

Fu, C., Samet, H., & Sankaranayananan, J. (2014). WeiboStand : Capturing Chinese Breaking News Using Weibo " Tweets ." In *7th ACM SIGSPATIAL Workshop on Location-Based Social Networks (LBSN'14)* (pp. 1–8). Dallas, TX.

Gabrielli, L., Rinzivillo, S., Ronzano, F., & Villatoro, D. (2014). From Tweets to Semantic Trajectories: Mining Anomalous Urban Mobility Patterns. In J. Nin & D. Villatoro (Eds.), *Citizen in Sensor Networks* (Vol. 8313, pp. 26–35). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-04178-0_3

Gao, S., Liu, Y., Wang, Y., & Ma, X. (2013). Discovering spatial interaction communities from mobile phone data. *Transactions in GIS*, *17*(3), 463–481. https://doi.org/10.1111/tgis.12042

Gao, S., Wang, Y., Gao, Y., & Liu, Y. (2013). Understanding Urban Traffic-Flow Characteristics: A Rethinking of Betweenness Centrality. *Environment and Planning B: Planning and Design*, *40*(1), 135–153. https://doi.org/10.1068/b38141

Geoghegan, J., Pritchard, L. J., Ogneva-Himmelberger, Y., Chowdhury, R. R.,

Sanderson, S., & Turner, B. L. (1998). “Socializing the Pixel” and “Pixelizing the Social” in Land-Use and Land-Cover Change. In D. Liverman, E. Moran, R. Rindfuss, & P. Stern (Eds.), *People and Pixels: Linkging Remote Sensing and Social Science* (1st ed., pp. 51–69). Washington D.C.: National Academy Press.

Gimpel, K., Schneider, N., O’Connor, B., Das, D., Mills, D., Eisenstein, J., … Smith, N. A. (2011). Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Shortpapers*, (2), 42–47. https://doi.org/10.1.1.206.3224

Girvan, M., & Newman, M. E. J. (2002, June 11). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.122653799

Glaeser, E. L., & Kahn, M. E. (2010). The greenness of cities: Carbon dioxide emissions and urban development. *Journal of Urban Economics*, *67*(3), 404–418. https://doi.org/10.1016/j.jue.2009.11.006

Goers, R. (2013). Who 鈥s Checking in to Downtown Tampa ? *Planning*, *79*(6), 36–39.

Goetz, S. J., Smith, A. J., Jantz, C., Wright, R. K., Prince, S. D., Mazzacato, M. E., & Melchior, B. (2003). Monitoring and predicting urban land use change applications of multi-resolution multi-temporal satellite data. In *IGARSS 2003. 2003 IEEE International Geoscience and Remote Sensing Symposium. Proceedings (IEEE Cat. No.03CH37477)* (Vol. 3, pp. 1567–1569). IEEE. https://doi.org/10.1109/IGARSS.2003.1294177

Gollege, R. G., & Stimson, R. J. (1997). *Spatial behavior: A geographic perspective*. Guilford Press.

González, M. C., Hidalgo, C. a, & Barabási, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, *453*(7196), 779–82. https://doi.org/10.1038/nature06958

Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, *69*(4), 211–221. https://doi.org/10.1007/s10708-007-9111-y

Guo, D., Zhu, X., Jin, H., Gao, P., & Andris, C. (2012). Discovering Spatial Patterns in Origin-Destination Mobility Data. *Transactions in GIS*, *16*(3), 411–429. https://doi.org/10.1111/j.1467-9671.2012.01344.x

Haklay, M. (2010). How Good is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets. *Environment and Planning B: Planning and Design*, *37*(4), 682–703. https://doi.org/10.1068/b35097

Hale, R., & Austin, D. M. (1997). An exploratory factor model of social area analysis. *Sociological Spectrum*, *17*(December), 115–125. https://doi.org/10.1080/02732173.1997.9982154

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques* (3rd ed.). Waltham, MA, USA: Morgan Kaufmann Publishsers.

Haralick, R., & Shapiro, L. (1985). Image segmentation techniques. *CVGIP: Image Understanding*, *29*(1), 100–132.

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., & Ratti, C. (2014). Geo-located Twitter as proxy for global mobility patterns. *Cartography and*

Geographic Information Science, *41*(3), 260–271. https://doi.org/10.1080/15230406.2014.890072

Hawley, A. H., & Duncan, O. D. (1957). Social Area Analysis: A Critical Appraisal. *Land Economics*, *33*(4), 337–345. Retrieved from http://www.jstor.org/stable/3144311

Hecht, B., & Stephens, M. (2014). A Tale of Cities: Urban Biases in Volunteered Geographic Information. In *ICWSM 14* (pp. 197–205). https://doi.org/papers3://publication/uuid/B13C63A5-B3B8-4619-9558-86BCAFE5E2CA

Herold, M., Couclelis, H., & Clarke, K. C. (2005). The role of spatial metrics in the analysis and modeling of urban land use change. *Computers, Environment and Urban Systems*, *29*(4), 369–399. https://doi.org/10.1016/j.compenvurbsys.2003.12.001

Heye, C., Leuthold, H., & Bourdieu, P. (2005). Theory-based social area analysis: an approach considering the conditions of a post-industrial society. *Area*, 1–7.

Ho, T. K. (1995). Random Decision Forests. In *Proceedings of the 3rd International Conference on Document Analysis and Recognition* (Vol. 47, pp. 278–282). Montreal, QC, Canada.

Hofmann, T. (1999). Probabilistic latent semantic indexing. *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '99*, 50–57. https://doi.org/10.1145/312624.312649

Hong, L., Fu, C., Torrens, P., Frias-Martinez, V., Fu, C., & Frias-Martinez, V. (2017). Understanding Citizens' and Local Governments' Digital Communications

During Natural Disasters. In *Proceedings of the 2017 ACM on Web Science Conference - WebSci '17* (pp. 141–150). New York, New York, USA: ACM Press. https://doi.org/10.1145/3091478.3091502

Hong, L., Torrens, P., Fu, C., & Frias-Martinez, V. (2017). Understanding citizens' and local governments' digital communications during natural disasters: The case of snowstorms. In *WebSci 2017 - Proceedings of the 2017 ACM Web Science Conference*. https://doi.org/10.1145/3091478.3091502

Hong, L., Yang, W., Resnik, P., & Frias-Martinez, V. (2016). Uncovering Topic Dynamics of Social Media and News: The Case of Ferguson. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 10046 LNCS, pp. 240–256). https://doi.org/10.1007/978-3-319-47880-7_15

Hu, B., & Ester, M. (2013). Spatial topic modeling in online social media for location recommendation. In *Proceedings of the 7th ACM conference on Recommender systems - RecSys '13* (pp. 25–32). New York, New York, USA: ACM Press. https://doi.org/10.1145/2507157.2507174

Hu, Y., Gao, S., Janowicz, K., Yu, B., Li, W., & Prasad, S. (2015). Extracting and understanding urban areas of interest using geotagged photos. *Computers, Environment and Urban Systems*, *54*, 240–254. https://doi.org/10.1016/j.compenvurbsys.2015.09.001

Huang, Q., & Wong, D. W. S. (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.

https://doi.org/10.1080/13658816.2016.1145225

Hussain, M., Chen, D., Cheng, A., Wei, H., & Stanley, D. (2013). Change detection from remotely sensed images: From pixel-based to object-based approaches. *ISPRS Journal of Photogrammetry and Remote Sensing*, *80*, 91–106. https://doi.org/10.1016/j.isprsjprs.2013.03.006

Iacono, M., Levinson, D., & El-Geneidy, A. (2008). Models of Transportation and Land Use Change: A Guide to the Territory. *Journal of Planning Literature*, *22*(4), 323–340. https://doi.org/10.1177/0885412207314010

Iceland, J., Weinberg, D. H., & Steinmetz, E. (2002). Appendix B. Measures of residential segregation. In *Racial and Ethnic Residential Segregation in the United States: 1980-2000*. Washington D.C.: U.S. Government Printing Office. Retrieved from https://www.census.gov/prod/2002pubs/censr-3.pdf

Inglada, J., & Christophe, E. (2009). The Orfeo Toolbox remote sensing image processing software. *IEEE International Geoscience and Remote Sensing Symposium*, (November), 733–736. https://doi.org/10.1109/IGARSS.2009.5417481

Intergovernmental Panel on Climate Change. (2014). Human Settlements, Infrastructure, and Spatial Planning. In *Climate Change 2014 Mitigation of Climate Change* (pp. 923–1000). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781107415416.018

Jackson, J. (2014). The Consequences of Gentrification for Racial Change in Washington, DC. *Housing Policy Debate*, *1482*(February), 1–21. https://doi.org/10.1080/10511482.2014.921221

Jacobs, J. (1961). *The Death and Life of Great American Cities* (1st ed.). New York, NY: Vintage Books, Random House.

Janson, C.-G. (1980). Factorial Social Ecology: An Attempt at Summary and Evaluation. *Annual Review of Sociology*, *6*(1 942), 433–456. https://doi.org/10.1146/annurev.so.06.080180.002245

Jenkins, A., Croitoru, A., Crooks, A. T., & Stefanidis, A. (2016). Crowdsourcing a collective sense of place. *PLoS ONE*, *11*(4), 1–20. https://doi.org/10.1371/journal.pone.0152932

Jia, Y., Ge, Y., Ling, F., Guo, X., Wang, J., Wang, L., … Li, X. (2018). Urban Land Use Mapping by Combining Remote Sensing Imagery and Mobile Phone Positioning Data. *Remote Sensing*, *10*(3), 446. https://doi.org/10.3390/rs10030446

Jiang, B. (2013). Head/Tail Breaks: A New Classification Scheme for Data with a Heavy-Tailed Distribution. *The Professional Geographer*, *65*(3), 482–494. https://doi.org/10.1080/00330124.2012.700499

Jiang, B., Ma, D., Yin, J., & Sandberg, M. (2016). Spatial Distribution of City Tweets and Their Densities. *Geographical Analysis*, *48*(3), 337–351. https://doi.org/10.1111/gean.12096

Johnson, I. L., Sengupta, S., Schöning, J., & Hecht, B. (2016). The Geography and Importance of Localness in Geotagged Social Media. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 515–526. https://doi.org/10.1145/2858036.2858122

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation

and Model Selection. *International Joint Conference on Artificial Intelligence*, *14*, 1137–1143. https://doi.org/10.1067/mod.2000.109031

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, *29*(1), 1–27. https://doi.org/10.1007/BF02289565

Kwan, M.-P. (1999). Gender, the home-work link, and space-time patterns of nonemployment activities. *Economic Geography*, *75*(4), 370–394. https://doi.org/10.1111/j.1944-8287.1999.tb00126.x

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159–174.

Lansley, G., & Longley, P. A. (2016). The geography of Twitter topics in London. *Computers, Environment and Urban Systems*, *58*, 85–96. https://doi.org/10.1016/j.compenvurbsys.2016.04.002

Lee, R., & Sumiya, K. (2010). Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection. *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks - LBSN '10*, 1. https://doi.org/10.1145/1867699.1867701

Lee, R., Wakamiya, S., & Sumiya, K. (2012). Urban area characterization based on crowd behavioral lifelogs over Twitter. *Personal and Ubiquitous Computing*, *17*(4), 605–620. https://doi.org/10.1007/s00779-012-0510-9

Leo, Y., Fleury, E., Alvarez-Hamelin, J. I., Sarraute, C., & Karsai, M. (2016). Socioeconomic correlations and stratification in social-communication networks. *Journal of The Royal Society Interface*, *13*(125), 20160598.

https://doi.org/10.1098/rsif.2016.0598

Li, L., Goodchild, M. F., & Xu, B. (2013). Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *Cartography and Geographic Information Science*, *40*(2), 61–77. https://doi.org/10.1080/15230406.2013.777139

Li, X., Zhang, C., & Li, W. (2017). Building block level urban land-use information retrieval based on Google Street View images. *GIScience & Remote Sensing*, *0*(0), 1–17. https://doi.org/10.1080/15481603.2017.1338389

Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, *37*(1), 145–151. https://doi.org/10.1109/18.61115

Lin, J., Benisch, M., Sadeh, N., Niu, J., Hong, J., Lu, B., & Guo, S. (2013). A comparative study of location-sharing privacy preferences in the United States and China. *Personal and Ubiquitous Computing*, *17*(4), 697–711. https://doi.org/10.1007/s00779-012-0610-6

Liu, Y., Kang, C., Gao, S., Xiao, Y., & Tian, Y. (2012). Understanding intra-urban trip patterns from taxi trajectory data. *Journal of Geographical Systems*, *14*(4), 463–483. https://doi.org/10.1007/s10109-012-0166-z

Liu, Y., Liu, X., Gao, S., Gong, L., Kang, C., Zhi, Y., … Shi, L. (2015). Social Sensing: A New Approach to Understanding Our Socioeconomic Environments. *Annals of the Association of American Geographers*, (April), 1–19. https://doi.org/10.1080/00045608.2015.1018773

Liu, Z. (2014). Applying a Spatio-Temporal Approach to the Study of Urban Social Landscapes in Tianjin , China.

Liu, Z. W., & Cao, H. H. (2017). Spatio-temporal urban social landscape transformation in pre-new-urbanization era of Tianjin, China. *Environment and Planning B-Urban Analytics and City Science*, *44*(3), 398–424. https://doi.org/10.1177/0265813516637606

Longley, P. A., & Adnan, M. (2016). Geo-temporal Twitter demographics. *International Journal of Geographical Information Science*, *30*(2), 369–389. https://doi.org/10.1080/13658816.2015.1089441

Loper, E., & Bird, S. (2006). NLTK: The Natural Language Toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions* - (pp. 69–72). Morristown, NJ, USA: Association for Computational Linguistics. https://doi.org/10.3115/1225403.1225421

Maat, K. (2009). *Built environment and car travel : analyses of interdependencies*. Amsterdam, The Netherland: IOS Press.

MacEachren, A. M. (2017). Leveraging Big (Geo) Data with (Geo) Visual Analytics: Place as the Next Frontier. In C. Zhou, F. Su, F. Harvey, & J. Xu (Eds.), *Spatial Data Handling in Big Data Era* (pp. 139–155). Springer, Singapore. https://doi.org/10.1007/978-981-10-4424-3_10

Malik, M. M., Lamba, H., Nakos, C., & Pfeffer, J. (2015). Population Bias in Geotagged Tweets. *9th International AAAI Conference on Weblogs and Social Media*, 18–27.

Maryland Department of Planning. (2010). 2010 Maryland Land Use Land Cover Map. Retrieved June 20, 2010, from http://mdpgis.mdp.state.md.us/landuse/imap/index.html

Massey, D. S., & Denton, N. A. (1988). The Dimensions of Residential Segregation. *Social Forces*, *67*(2), 281–315.

McKenzie, G., Adams, B., & Janowicz, K. (2015). Of Oxen and Birds: Is Yik Yak a Useful New Data Source in the Geosocial Zoo or Just Another Twitter? *Proceedings of the 8th ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, 4:1--4:4. https://doi.org/10.1145/2830657.2830659

McKenzie, G., & Janowicz, K. (2014). Coerced Geographic Information: The Not-so-voluntary Side of User-generated Geo-content. In *Proceedings of the 8th International Conference on geographic information science* (pp. 231–233). Vienna, Austria.

McKenzie, G., Janowicz, K., Gao, S., & Gong, L. (2015). How where is when? On the regional variability and resolution of geosocial temporal signatures for points of interest. *Computers, Environment and Urban Systems*, *54*, 336–346. https://doi.org/10.1016/j.compenvurbsys.2015.10.002

Mehrotra, R., Sanner, S., Buntine, W., & Xie, L. (2013). Improving LDA Topic Models for Microblogs via Tweet Pooling and Automatic Labeling. *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 889–892. https://doi.org/10.1145/2484028.2484166

Merler, S., & Ajelli, M. (2010). The role of population heterogeneity and human mobility in the spread of pandemic influenza. *Procedia Computer Science*, *1*(1), 2237–2244. https://doi.org/10.1098/rspb.2009.1605

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose.

In *International AAAI Conference on Weblogs and Social Media (ICWSM)* (pp. 400–408). Boston, MA.

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, *103*(23), 8577–8582. https://doi.org/10.1073/pnas.0601602103

Nishi, K., Tsubouchi, K., & Shimosaka, M. (2014a). Extracting Land-Use Patterns using Location Data from Smartphones. In *The 1st International Conference on IoT in Urban Space* (pp. 1–6). Rome, Italy. https://doi.org/10.4108/icst.urb-iot.2014.257220

Nishi, K., Tsubouchi, K., & Shimosaka, M. (2014b). Hourly pedestrian population trends estimation using location data from smartphones dealing with temporal and spatial sparsity. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - SIGSPATIAL '14* (pp. 281–290). New York, New York, USA: ACM Press. https://doi.org/10.1145/2666310.2666391

Oldenburg, R., & Brissett, D. (1982). The third place. *Qualitative Sociology*, *5*(4), 265–284. https://doi.org/10.1007/BF00986754

Olofsson, P., Foody, G. M., Herold, M., Stehman, S. V., Woodcock, C. E., & Wulder, M. A. (2014). Good practices for estimating area and assessing accuracy of land change. *Remote Sensing of Environment*, *148*, 42–57. https://doi.org/10.1016/j.rse.2014.02.015

Olofsson, P., Foody, G. M., Stehman, S. V., & Woodcock, C. E. (2013). Making better use of accuracy data in land change studies: Estimating accuracy and area and

quantifying uncertainty using stratified estimation. *Remote Sensing of Environment*, *129*(February), 122–131. https://doi.org/10.1016/j.rse.2012.10.031

Openshaw, S. (1984). *The Modifiable Areal Unit Problem*. Norwich, England: Geobooks.

Oshiro, T. M., Perez, P. S., & Baranauskas, J. A. (2012). How Many Trees in a Random Forest? In *Machine Learning and Data Mining in Pattern Recognition* (Vol. 7376, pp. 154–168). https://doi.org/10.1007/978-3-642-31537-4_13

Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). The PageRank citation ranking: Bringing order to the Web. In *Proceedings of the 7th International World Wide Web Conference* (pp. 161–172). Brisbane, Australia. https://doi.org/10.1.1.31.1768

Pan, G., Qi, G., Wu, Z., Zhang, D., & Li, S. (2013). Land-use classification using taxi GPS traces. *IEEE Transactions on Intelligent Transportation Systems*, *14*(1), 113–123. https://doi.org/10.1109/TITS.2012.2209201

Pappalardo, L., Pedreschi, D., Smoreda, Z., & Giannotti, F. (2015). Using big data to study the link between human mobility and socio-economic development. *2015 IEEE International Conference on Big Data (Big Data)*, (October), 871–878. https://doi.org/10.1109/BigData.2015.7363835

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pei, T., Sobolevsky, S., Ratti, C., Shaw, S.-L., Li, T., & Zhou, C. (2014). A new insight into land use classification based on aggregated mobile phone data. *International*

Journal of Geographical Information Science, *28*(October), 1988–2007. Computers and Society. https://doi.org/10.1080/13658816.2014.913794

Pettigrew, T. F. (2008). Future directions for intergroup contact theory and research. *International Journal of Intercultural Relations*, *32*(3), 187–199. https://doi.org/10.1016/j.ijintrel.2007.12.002

Puniyani, K., Eisenstein, J., Cohen, S. B., & Xing, E. (2010). Social Links from Latent Topics in Microblogs. *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, (June), 19–20. Retrieved from http://www.aclweb.org/anthology/W/W10/W10-0510

Quercia, D., Kosinski, M., Stillwell, D., & Crowcroft, J. (2011). Our Twitter Profiles, Our Selves: Predicting Personality with Twitter. In *2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Int'l Conference on Social Computing* (pp. 180–185). IEEE. https://doi.org/10.1109/PASSAT/SocialCom.2011.26

Ramage, D., Hall, D., Nallapati, R., & Manning, C. D. (2009). Labeled LDA. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing Volume 1 - EMNLP '09* (Vol. 1, p. 248). Morristown, NJ, USA: Association for Computational Linguistics. https://doi.org/10.3115/1699510.1699543

Ratti, C., Pulselli, R. M., Williams, S., & Frenchman, D. (2006). Mobile Landscapes: using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design*, *33*(5), 727–748. https://doi.org/10.1068/b32047

Reades, J., Calabrese, F., & Ratti, C. (2009). Eigenplaces: analysing cities using the space – time structure of the mobile phone network. *Environment and Planning B: Planning and Design*, *36*(5), 824–836. https://doi.org/10.1068/b34133t

Řehůřek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (pp. 45–50). Valletta, Malta: ELRA. https://doi.org/10.13140/2.1.2393.1847

Rodriguez Lopez, J. M., Heider, K., & Scheffran, J. (2017). Frontiers of urbanization: Identifying and explaining urbanization hot spots in the south of Mexico City using human and remote sensing. *Applied Geography*, *79*, 1–10. https://doi.org/10.1016/j.apgeog.2016.12.001

Rosenberg, A., & Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language (EMNLP-CoNLL'07)*, *1*(June), 410–420. https://doi.org/10.7916/D80V8N84

Saker, M., & Evans, L. (2016). Everyday life and locative play: an exploration of Foursquare and playful engagements with space and place. *Media, Culture & Society*, *38*(8), 1169–1183. https://doi.org/10.1177/0163443716643149

Sankaranarayanan, J., Samet, H., Teitler, B. E., Lieberman, M. D., & Sperling, J. (2009). TwitterStand. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS '09* (p. 42). New York, New York, USA: ACM Press.

https://doi.org/10.1145/1653771.1653781

Šćepanović, S., Mishkovski, I., Hui, P., Nurminen, J. K., & Ylä-Jääski, A. (2015). Mobile phone call data as a regional socio-economic proxy indicator. *PLoS ONE*, *10*(4), 1–15. https://doi.org/10.1371/journal.pone.0124160

Seto, K. C., Guneralp, B., & Hutyra, L. R. (2012). Global forecasts of urban expansion to 2030 and direct impacts on biodiversity and carbon pools. *Proceedings of the National Academy of Sciences*, *109*(40), 16083–16088. https://doi.org/10.1073/pnas.1211658109

Sexton, J. O., Song, X.-P., Huang, C., Channan, S., Baker, M. E., & Townshend, J. R. (2013). Urban growth of the Washington, D.C.–Baltimore, MD metropolitan region from 1984 to 2010 by annual, Landsat-based estimates of impervious cover. *Remote Sensing of Environment*, *129*, 42–53. https://doi.org/10.1016/j.rse.2012.10.025

Shevky, E., & Bell, W. (1955). *Social area analysis: theory, illustrative application, and computational procedures*. Stanford University Press.

Smith-Clarke, C., Mashhadi, A., & Capra, L. (2014). Poverty on the Cheap: Estimating Poverty Maps Using Aggregated Mobile Communication Networks. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 511–520. https://doi.org/10.1145/2556288.2557358

Song, C., Qu, Z., Blumm, N., & Barabási, A.-L. (2010). Limits of predictability in human mobility. *Science (New York, N.Y.)*, *327*(5968), 1018–21. https://doi.org/10.1126/science.1177170

Song, X. P., Sexton, J. O., Huang, C., Channan, S., & Townshend, J. R. (2016).

Characterizing the magnitude, timing and duration of urban growth from time series of Landsat-based estimates of impervious cover. *Remote Sensing of Environment*, *175*, 1–13. https://doi.org/10.1016/j.rse.2015.12.027

Soto, V., & Frias-Martinez, E. (2011). Robust land use characterization of urban landscapes using cell phone data. In *Proceedings of the 1st Workshop on Pervasive Urban Applications, in conjunction with 9th Int. Conf. Pervasive Computing* (pp. 1–8). https://doi.org/10.1.1.207.6031

Soto, V., Frias-Martinez, V., Virseda, J., & Frias-Martinez, E. (2011). Prediction of Socioeconomic Levels Using Cell Phone Records. In *International Conference on User Modeling, Adaptation, and Personalization* (pp. 377–388). https://doi.org/10.1007/978-3-642-22362-4_35

Spielman, S. E., & Thill, J. C. (2008). Social area analysis, data mining, and GIS. *Computers, Environment and Urban Systems*, *32*(2), 110–122. https://doi.org/10.1016/j.compenvurbsys.2007.11.004

Stefanidis, A., Crooks, A., & Radzikowski, J. (2011). Harvesting ambient geospatial information from social media feeds. *GeoJournal*, *78*(2), 319–338. https://doi.org/10.1007/s10708-011-9438-2

Tate, L. M., Suarez, S., Akundi, K., Pamela, Z., & Koempel, W. (2007). *The MD-355 / I-270 Technology Corridor Montgomery County , Maryland*. Retrieved from http://www.montgomeryplanning.org/research/documents/MD355I270web.pdf

Tobler, W. R. (1970). A Computer Movie Simulation Urban Growth in Detroit Region. *Economic Geography*, *46*, 234–240. https://doi.org/10.1126/science.11.277.620

Toole, J. L., Ulm, M., González, M. C., & Bauer, D. (2012). Inferring land use from

mobile phone activity. In *Proceedings of the ACM SIGKDD International Workshop on Urban Computing - UrbComp '12* (pp. 1–8). New York, New York, USA: ACM Press. https://doi.org/10.1145/2346496.2346498

Tsou, M.-H., Yang, J.-A., Lusher, D., Han, S., Spitzberg, B., Gawron, J. M., … An, L. (2013). Mapping social activities and concepts with social media (Twitter) and web search engines (Yahoo and Bing): a case study in 2012 US Presidential Election. *Cartography and Geographic Information Science*, *40*(4), 337–348. https://doi.org/10.1080/15230406.2013.799738

Tuan, Y. F. (1979). Space and Place: Humanistic Perspective. In S. Gale & G. Olsson (Eds.), *Philosophy in Geography* (pp. 387–427). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-009-9394-5

United Nations. (2014). *World Urbanization Prospects: The 2014 Revision, Highlights (ST/ESA/SER.A/352)*. *New York, United*. https://doi.org/10.4054/DemRes.2005.12.9

Van Arsdol, M., Camilleri, S., & Schmid, C. (1958). The Generality of Urban Social Area Indexes. *American Sociological Review*, *23*(3), 277–284.

Vázquez, A., Oliveira, J., Dezsö, Z., Goh, K., Kondor, I., & Barabási, A. (2006). Modeling bursts and heavy tails in human dynamics. *Physical Review E*, *73*(3), 36127. https://doi.org/10.1103/PhysRevE.73.036127

Voorhees, E. M. (1999). The TREC-8 Question Answering Track Report. *Natural Language Engineering*. https://doi.org/10.1017/S1351324901002789

Waddell, P., Wang, L., Charlton, B., & Olsen, A. (2010). Microsimulating parcel-level land use and activity-based travel: Development of a prototype application in San

Francisco. *Journal of Transport and Land Use*, *3*(2). https://doi.org/10.5198/jtlu.v3i2.124

Wakamiya, S., Lee, R., & Sumiya, K. (2011). Urban Area Characterization Based on Semantics of Crowd Activities in Twitter. In C. Claramunt, S. Levashkin, & M. Bertolotto (Eds.), *GeoS'11 Proceedings of the 4th international conference on GeoSpatial semantics* (Vol. 6631, pp. 108–123). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-20630-7

Walter, V. (2004). Object-based classification of remote sensing data for change detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, *58*(3–4), 225–238. https://doi.org/10.1016/j.isprsjprs.2003.09.007

Wang, W., & Stewart, K. (2015). Spatiotemporal and semantic information extraction from Web news reports about natural hazards. *Computers, Environment and Urban Systems*, *50*, 30–40. https://doi.org/10.1016/j.compenvurbsys.2014.11.001

Wang, Y., Wang, T., Tsou, M.-H., Li, H., Jiang, W., & Guo, F. (2016). Mapping Dynamic Urban Land Use Patterns with Crowdsourced Geo-Tagged Social Media (Sina-Weibo) and Commercial Points of Interest Collections in Beijing, China. *Sustainability*, *8*(11), 1202. https://doi.org/10.3390/su8111202

Widener, M. J., Farber, S., Neutens, T., & Horner, M. (2015). Spatiotemporal accessibility to supermarkets using public transit: An interaction potential approach in Cincinnati, Ohio. *Journal of Transport Geography*, *42*(July), 72–83. https://doi.org/10.1016/j.jtrangeo.2014.11.004

Wolf, J., Guensler, R., & Bachman, W. (2001). Elimination of the Travel Diary: Experiment to Derive Trip Purpose from Global Positioning System Travel Data.

Transportation Research Record: Journal of the Transportation Research Board, *1768*(1), 125–134. https://doi.org/10.3141/1768-15

Wu, L., Zhi, Y., Sui, Z., & Liu, Y. (2014). Intra-Urban Human Mobility and Activity Transition: Evidence from Social Media Check-In Data. *PLoS ONE*, *9*(5), e97010. https://doi.org/10.1371/journal.pone.0097010

Xian, G., Homer, C., & Fry, J. (2009). Updating the 2001 National Land Cover Database land cover classification to 2006 by using Landsat imagery change detection methods. *Remote Sensing of Environment*, *113*(6), 1133–1147. https://doi.org/10.1016/j.rse.2009.02.004

Xu, Y., Shaw, S. L., Zhao, Z., Yin, L., Lu, F., Chen, J., … Li, Q. (2016). Another tale of two cities: Understanding human activity space using actively tracked cellphone location data. *Annals of the American Association of Geographers*, *106*(2), 489–502. https://doi.org/10.1080/00045608.2015.1120147

Yang, J., Tsou, M., Jung, C., Allen, C., Spitzberg, B. H., Gawron, J. M., & Han, S. (2016). Social media analytics and research testbed (SMART): Exploring spatiotemporal patterns of human dynamics with geo-targeted social media messages, (June), 1–19. https://doi.org/10.1177/2053951716652914

Ye, M., Janowicz, K., Mülligann, C., & Lee, W. (2011). What you are is When you are: The Temporal Dimension of Feature Types in Location-based Social Networks. *Sigspatial*, 102. https://doi.org/10.1145/2093973.2093989

Yuan, J., Zheng, Y., & Xie, X. (2012). Discovering regions of different functions in a city using human mobility and POIs. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '12* (p.

186). New York, New York, USA: ACM Press. https://doi.org/10.1145/2339530.2339561

Zandbergen, P. a. (2009). Accuracy of iPhone locations: A comparison of assisted GPS, WiFi and cellular positioning. *Transactions in GIS*, *13*, 5–25. https://doi.org/10.1111/j.1467-9671.2009.01152.x

Zhao, B., & Sui, D. Z. (2017). True lies in geospatial big data: detecting location spoofing in social media. *Annals of GIS*, *23*(1), 1–14. https://doi.org/10.1080/19475683.2017.1280536

Zhao, W., Chen, J. J., Perkins, R., Liu, Z., Ge, W., Ding, Y., & Zou, W. (2015). A heuristic approach to determine an appropriate number of topics in topic modeling. *BMC Bioinformatics*, *16*(Suppl 13), S8. https://doi.org/10.1186/1471-2105-16-S13-S8

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E., Yan, H., & Li, X. (2011). Comparing Twitter and traditional media using topic models. *Advances in Information Retrieval*, 338–349. https://doi.org/10.1007/978-3-642-20161-5_34

Zhao, Z., Shaw, S. L., Xu, Y., Lu, F., Chen, J., & Yin, L. (2016). Understanding the bias of call detail records in human mobility research. *International Journal of Geographical Information Science*, *30*(9), 1738–1762. https://doi.org/10.1080/13658816.2015.1137298

Zhong, C., Arisona, S. M., Huang, X., Batty, M., & Schmitt, G. (2014). Detecting the dynamics of urban structure through spatial network analysis. *International Journal of Geographical Information Science*, *28*(11), 2178–2199. https://doi.org/10.1080/13658816.2014.914521

Zhou, T., & Li, H. (2014). Understanding mobile SNS continuance usage in China from the perspectives of social influence and privacy concern. *Computers in Human Behavior*, *37*, 283–289. https://doi.org/10.1016/j.chb.2014.05.008

Zhou, X., & Zhang, L. (2016). Crowdsourcing functions of the living city from Twitter and Foursquare data. *Cartography and Geographic Information Science*, *406*(June), 1–12. https://doi.org/10.1080/15230406.2015.1128852

Zook, M., Graham, M., Shelton, T., & Gorman, S. (2010). Volunteered Geographic Information and Crowdsourcing Disaster Relief: A Case Study of the Haitian Earthquake. *World Medical & Health Policy*, *2*(2), 6–32. https://doi.org/10.2202/1948-4682.1069