Quantifying Human Mobility Patterns During Disruptive Events With Geospatial Big Data

Yuqin Jiang

Follow this and additional works at: https://scholarcommons.sc.edu/etd

Part of the Geography Commons

Recommended Citation

This Open Access Dissertation is brought to you by Scholar Commons. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of Scholar Commons. For more information, please contact digres@mailbox.sc.edu.
QUANTIFYING HUMAN MOBILITY PATTERNS DURING DISRUPTIVE EVENTS
WITH GEOSPATIAL BIG DATA

by

Yuqin Jiang

Bachelor of Art
University at Buffalo, the State University of New York, 2014

Master of Science
University of South Carolina, 2017

Submitted in Partial Fulfillment of the Requirements
For the Degree of Doctor of Philosophy in
Geography
College of Arts and Sciences
University of South Carolina
2022

Accepted by:
Zhenlong Li, Major Professor
Susan L. Cutter, Committee Member
Michael E. Hodgson, Committee Member
Qunying Huang, Committee Member
Tracey L. Weldon, Vice Provost and Dean of the Graduate School
DEDICATION

I dedicate this work to my husband Andrey A. Popov, my father Jian Jiang, and my mother Mei Shen.
ACKNOWLEDGEMENTS

First, I want to thank my advisor, Dr. Zhenlong Li for giving me an opportunity to make this become real. Thank you for your trust, support, and challenges. I also want to thank my committee members, Dr. Susan L. Cutter, Dr. Michael E. Hodgson, and Dr. Qunying Huang, for your timely feedback and support.

Secondly, I want to thank the Department of Geography, University of South Carolina. I appreciate every professor and graduate students for your support, cheers, and this family-like environment. It takes a village to raise a child.

Thirdly, I want to thank my friends, especially Wang Hanyan and Zixuan Guo, for your supports in the last 15 years.

Last but not least, I want to thank my family, my husband Andrey A. Popov, my parents, my parents-in-law, and my cousin Yiwen Jiang. Thank you for your unconditional trust and support.
ABSTRACT

Understanding human mobility patterns is an essence for geography and geographical information science. Although existing studies have found that human mobility patterns are highly predictable, such patterns can be disrupted by events, ranging from sports games to natural hazard caused evacuations. However, traditional data collection methods that heavily rely on self-reported travel behaviors are often delayed and at a small scale, and thus are often not sufficient to reveal the disrupted human mobility patterns. Fortunately, with the development of geolocating-related technologies, multiple platforms are able to capture human mobility data in unprecedented spatiotemporal scales and granularities. These data, such as geotagged social media posts and vehicle travel records, are geospatial and Big in nature, characterized by “3Vs”: volume, variety, and velocity in the Big Data era. However, the quantitative methods used to analyze geospatial big data are lagging behind.

This dissertation contributes to the quantitative methodological developments in human mobility research with geospatial big data. Specifically, this dissertation responds to these three research questions: (1) how to detect spatiotemporal events using human mobility origin-destination data in an urban area? This research question aims to mine human mobility patterns from large volume data; (2) how did COVID-19 affect human mobility patterns over different land use types? This research question compares different data sources and aims to understand how human mobility patterns from various data sources differ; (3) how can aggregated long-term social connections help improve
evacuation modeling? This research question leverages the near real-time high velocity data and long-term historical data to understand evacuation behavior. This dissertation not only develops new methods to understand human mobility patterns under a specific event, but also attempts to advance quantitative methodologies that are used to analyze human mobility patterns with geospatial big data. Results of this dissertation demonstrate the potential in mining valuable information from geospatial big data by leveraging various quantitative methods and data sources. Methods developed in this dissertation can be applied to different applications and study areas with various data sources for future research.
# TABLE OF CONTENTS

Dedication ................................................................................................................................. iii

Acknowledgements ....................................................................................................................... iv

Abstract .......................................................................................................................................... v

List of Tables .................................................................................................................................. viii

List of Figures ................................................................................................................................. ix

Chapter 1: Introduction .................................................................................................................. 1

Chapter 2: A Novel Method for Spatiotemporal Event Detection Using Discrete
Empirical Interpolation Method ...................................................................................................... 7

Chapter 3: Spatiotemporal Patterns of Human Mobility and Its Association with Land
Use Types During COVID-19 In New York City ........................................................................ 35

Chapter 4: Social Distance Integrated Gravity Model for Evacuation Destination
Choice ............................................................................................................................................ 61

Chapter 5: Conclusion ..................................................................................................................... 85

References ......................................................................................................................................... 90
LIST OF TABLES

Table 2.1 Locations of the 13 sensors for 2009 ......................................................... 25
Table 2.2 Locations of the 8 sensors for 2012 ............................................................. 30
LIST OF FIGURES

Figure 2.1 Illustration of the overall workflow ................................................................. 14

Figure 2.1 Modeling the spatiotemporal human mobility data. (a) Space-time
cube in the geographic space summarizing the human mobility data,
and (b) Matrix A transformed from the space-time cube ................................. 15

Figure 2.3 Determine sensors locations ................................................................. 17

Figure 2.4 Workflow of finding the optimal number of sensors and their locations .... 19

Figure 2.5 Travel demand by taxi in NYC of January 2009 ........................................ 22

Figure 2.6 Determination of the optimal number of sensors for 2009(a) and 2012(b) .... 24

Figure 2.7 Locations of 13 sensors in NYC for 2009 event detection ....................... 25

Figure 2.8 Daily RMSE showing the difference between observed travel
demand vs simulated uneventful scenario for 2009 ........................................... 26

Figure 2.9 Spatial distribution of Anomaly Index for Thanksgiving Day 2009 ....... 27

Figure 2.10 Anomaly Index of Manhattan area in 2009 for (a) June 28th
(Pride March celebrating the LGBTQ community), (b) Nov. 1st
(NYC Marathon), and (c) Dec. 25th (Christmas) ............................................. 28

Figure 2.11 Locations of the 8 sensors for 2012 ........................................................ 30

Figure 2.12 Daily RMSE showing the difference between observed travel
demand vs simulated uneventful scenario for 2012 ....................................... 31

Figure 2.13 Anomaly Index of Manhattan area in 2012 for (a) Nov. 23rd
(Thanksgiving Day), (b) Oct. 29th (Hurricane Sandy) (c) Mar. 18th
(St. Patrick’s Day), and (d) Dec. 31st, 2012 (New Year’s Eve) ............................ 32

Figure 3.1 Human mobility changing patterns at parcel-level in week 1 – week 9 ....... 44

Figure 3.2 Human mobility changing patterns at parcel-level in week 10 – week 15 ..... 45

Figure 3.3 Mobility changing patterns for six land use types based on Twitter data ...... 46
Figure 3.4 Twitter- and Google-based human mobility changing patterns in Bronx County

Figure 3.5 Twitter- and Google-based human mobility changing patterns in Kings County

Figure 3.6 Twitter- and Google-based human mobility changing patterns in New York County

Figure 3.7 Twitter- and Google-based human mobility changing patterns in Queens County

Figure 4.1 Hurricane Matthew path and the 10 selected counties

Figure 4.2 Evacuation destination distribution

Figure 4.3 Evacuation destination popularity vs familiarity rank

Figure 4.4 The social connection for the four selected counties

Figure 4.5 Cross-validation process
CHAPTER 1
INTRODUCTION

“Human mobility” is a commonly used but loosely defined term. It represents the concept about peoples’ spatiotemporal occupation and involves interaction among human, society, and the surrounding physical environment. Human mobility studies human movement and displacement across space and time. Human mobility also conflates some other terms including human dynamics, human movements, and human motions (Shaw et al., 2016; M. Yuan, 2018).

Understanding human mobility is an essential part of geography and spatial science, as it investigates human interactions with the surrounding physical environment and the use of geographic space (Batty, 2008; Cresswell, 2011, 2012, 2014). Furthermore, goods, capitals, knowledge, diseases, and many other things also travel as humans travel (Barbosa et al., 2018). Since the “quantitative movements” in geography in the 1950s to 1960s, advances in human movement research have been observed theoretically and methodologically (Barnes, 2004; Burton, 1963; Harvey, 1969; Johnston, 1981). One major trend that emerged is to borrow concepts and models from physics about the movements of objects, and applied them to geography and spatial science to investigate human movements. For instance, Barabási (2005) mathematically proved that human activities occur in bursts and exhibit a heavy-tail distribution. Following this, a heavy-tail distributed power-law model emerged as one of the most commonly used distribution to model human travel distances across space (Brockmann et al., 2006;
Gonzalez et al., 2008; Mandelbrot & Mandelbrot, 1982). The most well-known example is the gravity model. Rooted in Newton’s law of gravity, the gravity model for human mobility considers the number of individuals’ movements between two places as “force” (Anderson, 2011; Simini et al., 2012). The gravity model is considered as a universal model to explain human mobility patterns as it fits in multiple applications, including migration (Karemera et al., 2000; Simini et al., 2012; X. N. Zhang et al., 2020), communications (Krings et al., 2009; Palchykov et al., 2015), transportation infrastructure (Hong & Jung, 2016; W.-S. Jung et al., 2008), and hazard-caused evacuation (G. Cheng et al., 2011; Y. Jiang, Li, et al., 2021; Wilmot et al., 2006).

Traditionally, data collection is the bottleneck of human mobility related research. Such data are often collected using active methods, meaning participants are aware that their information is collected for specific research purposes. For example, long-term migration decisions are reported by the Census Bureau at county-level, based on self-reported surveys (Guo & Zhu, 2014; Lichter & Johnson, 2006). Protective evacuations for an upcoming natural hazard are mainly collected by post-disaster surveys (Martín et al., 2017a, 2020). For short-term travel patterns, data are collected by small-scale observations or are self-reported (Gong et al., 2012; Rhee et al., 2011). However, these methods are delayed and expensive. Census reports reveal geographically aggregated human mobility patterns every five years (US Census Bureau, 2021). Large-scale surveys are both financially expensive and labor-intensive (Y. Jiang, Li, & Cutter, 2019; Martín et al., 2017a). In addition, whether survey respondents reported their actual travel behaviors or not cannot be validated (Martín et al., 2020). Small-scale observations may
result in biased data, which could cause unreliable conclusions (Greenland, 2005; Nieto et al., 2020).

Around 2010, an abstract concept called “Big Data” emerged and immediately caught massive research attention. “Big Data” is commonly characterized by the “3Vs:” Volume, Variety, and Velocity (M. Chen et al., 2014; Goodchild, 2016). Volume directly points out the massive size of data. Velocity refers to the rapidity in data generation and measurement of the timeliness of data. Variety indicates the numerous data types and multiple dimensions contained in the data. The speed of data generation has been consistently increasing, causing the size of data to grown exponentially (M. Chen et al., 2014). According to a report published by International Data Corporation, the global data size has reached 33 Zettabytes (1 Zettabyte = 10^9 Terabytes), and is predicted to be 175 Zettabytes by 2025 (Reinsel et al., 2018). While these 3Vs captured the characteristics of Big Data, a more important aspect of Big Data is its Value (Z. Li, 2020;).

The Big Data Era has significantly impacted GIScience in terms of data collection and analysis methods. With the prevalence of smart phones and location sensors (e.g., GPS), capturing, receiving and sharing geolocations has become effortless. The ability to capture and share real-time location from personal devices increased the velocity of geospatial data generation. This directly results in an increasing volume of geospatial data. An example is geotagged social media posts. When a post is associated with a location, it marks an individual’s geographic position at the given moment, which is enriched by text, images, videos, and other information. Besides social media posts, transportation data, cell phone communications, and many other data with geolocations
are considered as geospatial data. The increasing variety of geospatial data simultaneously increases its volume and velocity.

Geospatial big data illustrates its potential as a low-cost and less labor-intensive solution in data collection processes for scientific research and is emerging as a new data source to examine human mobility patterns. For example, during Hurricane Sandy, Twitter-based human mobility data were collected to understand the perturbation (Q. Wang & Taylor, 2014). During Hurricane Harvey, platforms were developed to collect real-time hurricane-related tweets for emergency management (Ahmouda et al., 2019; Alam et al., 2018; J. Yang et al., 2019). In addition, geospatial big data have played an important role in contact tracing and understanding the spread of infectious diseases (Heiler et al., 2020; Vazquez-Prokopec et al., 2013; Wesolowski et al., 2015).

With big human mobility data, studies have found that individuals’ mobility patterns show regularity during uneventful situations and are highly predictable (C. Lu et al., 2013; Song et al., 2010). At the same time, such regularity can be disrupted by events, especially from urban areas. Some events are planned and thus can be expected, such as sports games and festival parades (Giannotti et al., 2011; Pan et al., 2013a). People can expect road closures and massive gathering and thus can plan in advance. Some events can be expected but the intensity and impact areas may vary, such as hurricanes or winter storms (Martín et al., 2017a). Some other events are completely unpredictable, such as major car accidents (C. Chen et al., 2016). Those unexpected events have various impact scales and may not even be recorded. Geospatial big data have helped to record such events, and thus, have been utilized for understanding their human mobility patterns. For
instance, mobile phone location data and geotagged social media data illustrated changes of travel behaviors in the United States during the COVID-19 pandemic (S. Gao, Rao, Kang, Liang, Kruse, et al., 2020; Z. Li et al., 2021). Recent developments focused on how to collect, store, and visualize geospatial big data with massive volume, high velocity, and wide variety, but quantitative methods used to analyze those geospatial big data are lagging behind. Therefore, better and more robust quantitative methods to mine the value of geospatial big data are needed.

This dissertation contributes to quantitative methodological developments in human mobility research using geospatial big data. Specifically, this dissertation is organized with three research questions:

Chapter 2: How to detect spatiotemporal events in urban areas using human mobility data?

This chapter proposes an event detection method that handles large volumes of human mobility data. Specifically, this method applies the Discrete Empirical Interpolation Method to first identify important spatial locations in the study area and then simulates a regular uneventful scenario based on observation data from those important locations. By comparing the simulated scenario with the observation data, events can be detected spatially and temporally. This is a data-driven method that does not require arbitrary threshold choices or prior knowledge about the study area or time span.

Chapter 3: How did COVID-19 affect human mobility patterns in different land use types?
This chapter handles the variety aspect of big data by comparing various data sources. Specifically, this chapter compares human mobility pattern changes detected with geotagged Twitter data, an open access data source, and with the Google Community Mobility Report, a private data source with limited access. Comparison results show that Twitter data can reveal similar patterns to the Google Community Mobility Report. Results in this chapter validate the usefulness of open social media-based human mobility data sources.

Chapter 4: How can aggregated long-term social connections help improve evacuation modeling?

This chapter leverages near real-time evacuation-related data with long-term social media data to explore the potential of long-term social media data in understanding hurricane evacuation behavior. Specifically, this chapter first develops a big data approach to measure county-to-county social distance using long-term geotagged tweets, and then integrate this social distance to the traditional gravity model. This social distance integrated gravity model shows an improvement of 35% accuracy in predicting evacuation destinations comparing to the gravity model that only considered physical distance.
CHAPTER 2
A NOVEL METHOD FOR SPATIOTEMPORAL EVENT DETECTION USING DISCRETE EMPIRICAL INTERPOLATION METHOD

2.1 INTRODUCTION

Human movements in urban areas are the pulse of the city. Human movements indicate the functionality of a city and how people are interacting with their surrounding physical environment (Peng et al., 2012; J. Yuan et al., 2012). Understanding human mobility patterns in a city is essential for urban planning (Barbosa et al., 2018; Isaacman et al., 2012), transportation (C. Chen et al., 2016; Z. Huang et al., 2018), and emergency management (Y. Jiang, Li, & Cutter, 2019; Martín et al., 2017a). Fortunately, the availability of human mobility data offers researchers great opportunities, including cell phones records (X. Huang et al., 2021; X. Lu et al., 2012; Song et al., 2010), geo-located social media posts (X. Huang, Li, Jiang, Li, et al., 2020b; Y. Jiang, Huang, et al., 2021), and travel records (Y. Jiang, Guo, et al., 2021; X. Zhu & Guo, 2017). Researchers from multiple disciplines have contributed to better modeling and understanding of the regularity and anomaly of human mobility patterns.

Activities in the city are complicated and full of uncertainties. Events or anomalies are hard to define due to the complexity and scale of a city. Some events are

---

planned and can be expected, such as sports games and festival parades (Giannotti et al., 2011; Pan et al., 2013a). People can expect road closures and massive gatherings and thus, can plan in advance. Some events can be expected but the intensity and impact areas are uncertain, such as hurricanes or winter storms (Martín et al., 2017a). Some other events are unexpected, such as car accidents, blackouts, or earthquakes (Q. Chen et al., 2016). Those unexpected events have various impact scales and may not even be recorded. Urban event or anomaly detection methods have been developed with multiple data sources, including trajectory (Khezerlou et al., 2017; Piciarelli et al., 2008; H. Wu et al., 2017), origin-destination (OD) trip records (X. Li et al., 2009; Zheng et al., 2015; X. Zhu & Guo, 2017), social media posts (Dhiman & Toshniwal, 2020; Weng & Lee, 2011; X. Zhou & Chen, 2014), cell phone data (Dobra et al., 2015; Traag et al., 2011), and videos/images from surveillance cameras (Adam et al., 2008; C. Lu et al., 2013; Oh et al., 2011; Wan et al., 2020).

From the data type perspective, point-based data is most commonly used. A point in human mobility can be retrieved from a geotagged social media post, a trip origin or destination location, or a cell phone connected to a tower location. Existing spatiotemporal event detection methods for point-based data include statistical and probabilistic methods, supervised and unsupervised machine learning methods, and artificial intelligence methods (Toch et al., 2019; Yu et al., 2020; M. Zhang et al., 2020).

In this chapter, a novel big data method for event detection with spatiotemporal data was purposed. The method is able to identify important locations in the data and simulate a regular uneventful scenario. The discrepancy between simulation and all the observed data defines the events in space and time. Specifically, we make use of the
Discrete Empirical Interpolation Method (DEIM), which extends Principal Component Analysis (PCA) to decompose the data, not only into its principal components in the data space, but also key spatial locations in the geographic space (Brunton & Kutz, 2019; Štefănescu et al., 2015). We apply this method to billions of taxi OD trip records in New York City (NYC) during 2009. Results show that this method can identify the most anomalous days during the year and can display the spatial patterns of discrepancy.

2.2 RELATED WORK

Event detection or anomaly detection methods using human mobility data have been extensively studied and used in multiple areas, including city-scale event detection (Abdelhaq et al., 2013; Costa et al., 2018; Hu et al., 2017; W. Zhang et al., 2015), traffic conditions (Pan et al., 2013a; S. Zhang et al., 2015), environment management (W. Jiang et al., 2015; Resch et al., 2018; F. Zhang et al., 2019), infectious diseases (Y. Gao et al., 2018; Y. Jiang, Huang, et al., 2021; J. Wang et al., 2018), and natural hazards (Y. Jiang, Li, et al., 2021; Sakaki et al., 2010; Y. Wang & Taylor, 2018; Yu et al., 2019).

Those methods have been developed based on multiple data sources, including social media data (W. Jiang et al., 2015; Y. Jiang, Huang, et al., 2021; C. Zhang et al., 2016), vehicle trajectory data (Andrienko et al., 2011; Cui et al., 2016; Ying et al., 2014), OD trip records (Jahnke et al., 2017; Kaiser et al., 2017; X. Zhu & Guo, 2017), and cell phone data (Dobra et al., 2015; Fekih et al., 2020). Due to the different strengths and weaknesses of each data type, methods and applications are tailored to make the most use of each data source. For example, social media data contains rich information, other than geolocation, it usually includes text, user’s profile, and sometimes images or videos. Methods and tools have been developed to merge geolocation with topics retrieved from
text, images, or videos, to identify events or anomalies across space and time. For example, studies have applied topic modeling methods to identify the most dominant topics discussed on Twitter at the given space and time to identify events (Abd-Alrazaq et al., 2020; X. Chen et al., 2018; Jelodar et al., 2019). In addition, recent methodology development in image recognition has improved accuracy in event detection with social media data by fusing text and image (Alqhtani et al., 2015; X. Huang et al., 2019).

However, social media data suffers from the representativeness issue (Y. Jiang, Li, & Ye, 2019a; Malik et al., 2015a; Mellon & Prosser, 2017). Users of social media are biased, and the majority of social media posts do not contain geotags. Cell phone or mobile phone data have a better population representation and high penetration. However, the Call Detail Record (CDR) data, the widely used mobile phone data, is based on the locations of cellular towers. The data records the cell tower’s location when a cell phone user is making a call or sending/receiving a text message. Therefore, the geographic precision relies on the spatial density of signal towers (Resch et al., 2018; Traag et al., 2011; X. Zhu & Guo, 2017). Vehicle-based datasets show higher resolution, especially in urban areas. There are two types of vehicle-based mobility data. One is trajectory data and the other is origin-destination (OD) data. Trajectory data records the vehicle’s location at a certain time interval, which can indicate the actual route; however, the status of the vehicle is unknown, such as whether a taxicab is occupied or empty (Piciarelli et al., 2008; Toch et al., 2019; S. Zhou et al., 2015). OD data, on the other hand, only records a trip’s origin and destination locations, ignoring the actual driving route. Although the exact driving routes remain unknown, OD data marks the demand for travel with vehicles across space and time (Jahnke et al., 2017; Kaiser et al., 2017). Studies with
OD taxi data have been used to help understanding the landscapes of cities, to improve taxi dispatches, and to identify people’s travel behaviors (Peng et al., 2012; J. Tang et al., 2015; X. Zhu & Guo, 2017). In this paper, we use taxi trip OD data. More specifically, we treat each trip origin and destination as a point.

Point-based event detection methods include clustering and time-series analysis. Clustering methods are unsupervised learning approaches. For example, multiple clustering algorithms can partition a city into multiple functionality regions, and thus identify anomalies in human mobility patterns with OD trip data. For example, DBSCAN has been used to identify hotspots in OD trips as clusters for pick up and drop off activities (Jahnke et al., 2017; J. Tang et al., 2015). DBSCAN clusters can also identify activities with social media check in data (Luo et al., 2016). The K-means clustering algorithm has been used to partition a city into multiple regions based on taxi pick up and drop off activities (J. Tang et al., 2018). The K-means clustering method has also been used to identify regularities in an individual’s mobility pattern, and thus, can help to identify anomalies from outliers of clusters (Qin et al., 2012). In another study (J. Yuan et al., 2012), origin and destination trip records have been constructed into cubes and then used the Latent Dirichlet Allocation to identify clusters. Network-based clustering methods are used to detect significant OD flow based on sharing bike data (Zaltz Austwick et al., 2013). However, these clustering methods focus more on the spatial distribution patterns at a given time, and thus may ignore long-term trends (X. Zhu & Guo, 2017). To overcome this, studies have applied methods to detect temporal trends with time-series analysis. For example, Discrete Fourier Transformation has been used to detect periodicity in human mobility patterns, which can separate daily and weekly
regular mobility pattern changes from anomalies (W. Zhang et al., 2015). To further
decompose long-term and seasonal trends, a seasonal and trend decomposition method
has been applied, which can decompose time-series pattern into long-term trend, seasonal
periodicity, and the remainder, where the significant remainders are considered as events
(X. Zhu & Guo, 2017). However, time-series analysis methods require data to be sorted
chronologically, which require intensive computational power when data size is large.

In some fields when the complete observation dataset is not available, event or
anomaly detection relies on simulation with limited observation data points. Such
simulation-based event detection methods are used when observation data can be only
captured with a physical sensor in the field. For example, hydrological models rely on
observation data from gauges to estimate the probability for a flood depth (Alfieri et al.,
2012; Younis et al., 2008). Numerical weather prediction relies on data assimilation
methods whereby model forecasts are supplemented by sparse noisy observations of the
atmosphere in a Bayesian framework (Evensen, 2009; Kalnay, 2003). In this paper, we
propose a novel event or anomaly detection approach using DEIM (Brunton & Kutz,
2019; Ţefănescu et al., 2015), which is a simulation method based on PCA. PCA can
transform original dataset into orthonormal components, which indicate the correlation
between a new predictive variable and observation variables. Based on this concept, PCA
related algorithms have been developed to identify the most dominant factors in traffic
patterns, and thus, anomalies can be detected (Chawla et al., 2012; S. Yang et al., 2014;
S. Yang & Zhou, 2011). DEIM further extends PCA by identifying not only the most
dominant linear combination of variables, but also the most dominant variable in each
component. DEIM has been used in areas including nuclear desalination plant
(Upadhyaya & Li, 2011), water distribution (Liu & Auckenthaler, 2014), and fluids flow reconstruction (Jayaraman et al., 2019).

2.3 METHODOLOGY

2.3.1 Method Review

This section proposes an anomaly detection method with optimal sensor placement and interpolation using DEIM. This method consists of four steps. The first step is data preparation. In this step, we divide the study area into spatial cells and count travel activities in each cell. In the second step, we determine the optimal number of sensors and their locations. In this study, a sensor is an abstract concept such that only the true observation data at the sensor location will be treated as known data for simulation. The sensor’s observation data is equivalent to the number of travel activities in the cell to which the sensor belongs. The third step is to simulate a regular uneventful scenario based on the observation data at the sensors’ locations. The last step is to compare the simulated situations and the observation. At the aggregated level, we calculate Root-Mean-Squared-Error (RMSE) to identify which temporal unit has the largest anomalies. Differences at the cell-level can be mapped for the spatiotemporal distribution of the anomalies. Figure 2.1 shows the overall workflow in this section.
2.3.2 Model Human Mobility Data

We partition the study area into small spatial units, commonly rectangle or square shaped cells. Each cell is assigned with a unique identification number. Then, for each temporal unit (e.g., a day or an hour), we summarize the human mobility signals in each cell. Such signals can be the number of taxi pick-ups or drop-offs, the number of geotagged social media posts, or cell phone signals. By mapping human mobility data per cell, we can generate a cell-based 2-dimensional map for one temporal unit. Creating and overlaying such maps for all the temporal units, we create a space-time cube from human mobility data (Figure 2.2a). In this space-time cube, X and Y correspond to the location of the given cell in the 2-dimensional map. In this example, there are \( n \) rows and \( m \)
columns for the 2-dimensional map, and thus, there are total of $n \times m$ cells for each temporal unit. The depth of this cube, $T$, represents the number of temporal units.

We then transform this space-time cube into a matrix $A$ (Figure 2.2b). This matrix has $k$ columns, representing $k$ temporal units. The number of rows for this matrix is $n \times m$, representing a vectorized 2-dimensional map. The row identification number corresponds to the unique cell identification number in the study area. This matrix $A$ is used for later analysis.

2.3.3 Determine the Locations of Sensors

Locations of sensors are determined by PCA. PCA is a robust method to handle raw data where the number of samples is larger than the number of uncorrelated variables is relatively small (Good & Hardin, 2012; S. Jung & Marron, 2009). In other words, PCA is robust when the input matrix has more columns than rows. When the input dataset has more variables than samples (i.e., the input matrix has more rows than columns), PCA will generate spurious correlations, which does not reflect the true relationship.
Unfortunately, this is the common case in spatiotemporal analysis as we usually have more spatial units (rows) than temporal units (columns). To solve this problem, DEIM further extends PCA by not only identifying the dominant vector, but also identifying the dominant variables. In PCA, each component takes all the spatial units, which means each component needs data from all spatial units. DEIM takes the first few dominant components and identifies the most important spatial unit in each component. The goal of DEIM is to find the most dominant spatial location that indicates the strongest correlation. Specifically, DEIM sequentially looks at each of the dominant components from PCA and finds the variable with the largest coefficient.

In this demonstrative example, the original dataset matrix has \( k \) temporal units (\( k \) columns). Let \( x \) be the total spatial unit, where \( x = m \times n \) from Figure 2.2b. Therefore, the component matrix resulted from PCA has \( k \) components in total. The first \( q \) components explain 96\% of the variance and the remainder of the components \((k - q)\) explain 4\% of the variance (Figure 2.3a).

Let matrix \( U \) be the first \( q \) component (the blue rectangle). The size of \( U \) is \( x \) by \( q \), where \( x \) is the number of spatial units and \( q \) is the number of first \( q \) dominant components we decide to use. Since the number of spatial units is usually significantly larger than the number of components, matrix \( U \) is a long and narrow matrix. The transpose matrix of \( U \) is \( U^T \), which is a wide and short matrix.
Because \( \mathbf{U} \) is the matrix for dominant components, it is an orthonormal matrix, and so is \( \mathbf{U}^\top \). Therefore, \( \mathbf{U} \mathbf{U}^\top \) removes all the non-dominant components from the PCA result. \( \mathbf{U} \mathbf{U}^\top \) is about equal to an identity matrix, but not exactly the same. Because \( \mathbf{U} \mathbf{U}^\top \) removes all the non-dominant components, \( \mathbf{U} \mathbf{U}^\top \mathbf{A} \) creates an approximation of the original dataset \( \mathbf{A} \).

\[
\mathbf{A} \approx \mathbf{U} \mathbf{U}^\top \mathbf{A}
\]  
Eq. 2.1

In this example, matrix \( \mathbf{U} \) is for the first \( q \) components. Therefore, the size of \( \mathbf{U} \) is \( x \times q \). In each component (each column of \( \mathbf{U} \)), we find the element with the largest absolute value. The first dominant component represents the first sensor, the second component represents the second sensor. This process goes on until we reach the predetermined number of sensors. In this example, we will have \( q \) sensors. For example, the \( q \) red squares (spatial units) in figure 2.3 indicate locations for the \( q \) sensors.
After we determine the locations of the $q$ sensors, we can define a matrix $P$, representing the sensors’ locations. In $P$, cells in which a sensor is located will have a value 1, and all other cells have a value 0. The size of matrix $P$ is $x \times q$, same as $U$, but all elements of $P$ are 0, except elements at the $q$ locations of sensors which have a value 1 (Figure 2.3b). Because $P$ is also orthonormal, $P P^T$ is about equal to an identity matrix.

2.3.4 Determine the Optimal Number of Sensors

In this anomaly detection model, the only parameter is the number of sensors. To determine the optimal number of sensors to use, we first train the model using the dataset in which we are interested. However, because this method is based on PCA, the more sensors it takes, the less residual error will result, which will result in an overfitted model. To solve this, we bring an external validation dataset to find out the optimal number of sensors to use. Validation dataset is very similar to training dataset. For example, they can be from the same study area but during different time periods.

We first apply this model to the training dataset with different numbers of sensors. Each run uses different number of sensors and thus will generate a corresponding $RMSE$.

Let $a_{ij}$ be the element from $A$ at $i$th row and $j$th column and $\tilde{a}_{ij}$ be the element from $\tilde{A}$ at $i$th row and $j$th column. We calculate the $RMSE$ to compute the difference between observed dataset $A$ and the simulated data $\tilde{A}$ using the $RMSE$ calculation method:

$$RMSE = \sqrt{\frac{\sum(a_{ij} - \tilde{a}_{ij})^2}{n}}$$

where $n$ is the number of cells in the study area. This $RMSE$ represents how big the difference is between the observed dataset and simulated dataset.

$RMSE$ is a measurement for the overall simulation error for the dataset. Therefore, the training dataset and the validation dataset do not need the same number of temporal
units. In the validation round, we apply the locations of sensors to the validation dataset for simulation. The curve for models with the training dataset will continue decreasing as the number of sensors increases, but the \( RMSE \) for validation dataset will reach a lowest point. We use the number of sensors corresponding to the lowest \( RMSE \) point as the optimal number of sensors. This workflow is shown in Figure 2.4.

Figure 2.4 Workflow of finding the optimal number of sensors and their locations
2.3.5 Simulate the Uneventful Scenario

A traditional sensor is physical equipment put in the field to collect data. Based on the discrete data collected from sensors, the whole dataset is simulated. In our application, we treat sensors as the spatial unit with a known data point in all the temporal units to simulate a complete dataset. As we only use $q$ sensors from the first $q$ components, such simulation is an approximation of the original dataset. For example, in Figure 2.3a, the first $q$ components explain 96% of the variance of the dataset. We assume that this simulated uneventful situation ($\tilde{A}$) is about 96% time equal to the original dataset ($A$), and events are contained in the remaining 4% unexplained variance.

Since the only known data are at the sensors’ locations, these known data can be written as $P^T A$. Given Eq.2.1, we can get

$$P^T A \approx P^T U U^T A$$  \hspace{1cm} Eq.2.3

Then divide $P^T U$ from both sides of Eq.2.3, we get

$$(P^T U)^{-1} P^T A \approx U^T A$$  \hspace{1cm} Eq.2.4

Then multiple $U$ on both sides of Eq.2.4, we get

$$U(P^T U)^{-1} P^T A \approx U U^T A.$$  \hspace{1cm} Eq.2.5

Because $U U^T A$ is an approximation of the original dataset $A$ (Eq.1), we can rewrite Eq.2.5 into

$$\tilde{A} \approx U(P^T U)^{-1} P^T A$$  \hspace{1cm} Eq.2.6

Eq.2.6 is the equation used to simulate the scenario without anomaly.

2.3.6 Detect Events

The simulated uneventful scenario $\tilde{A}$ has the same size as the original dataset $A$. Specially, each row and column match the corresponding spatial and temporal unit. We
first calculate the difference between observation and simulation for each spatial unit and
temporal unit using the following Eq. 2.2: Let $a_{ij}$ be the element from $A$ at $i$th row and $j$th
column and $\tilde{a}_{ij}$ be the element from $\tilde{A}$ at $i$th row and $j$th column. We calculate the Root-
Mean-Squared-Error (RMSE) to compute the difference between observed dataset $A$ and
the simulated data $\tilde{A}$ using the RMSE calculation method (Eq. 2.2).
where $n$ is the number of cells in the study area. This RMSE represents how big the
difference is between observed dataset and simulated dataset.

For a given temporal unit, we pair up the observed travel demand and the
simulated travel demand and compare the difference. We define the difference as
“Anomaly Index” by using the observation travel demand number minus the simulated
travel demand number (Eq. 2.7):

$$\text{Anomaly Index} = \text{observed} - \text{simulated} \quad \text{Eq. 2.7}$$

We then map the Anomaly Index for each spatial unit to identify the spatial distribution
of anomalies.

2.4 CASE STUDY

New York City (NYC) consists of five boroughs: Brooklyn, Queens, Manhattan,
the Bronx, and Staten Island. Based on the 2010 Census data, NYC has more than 8
million residents living in about 800 km$^2$, which means NYC has the highest population
density in the United States. Among these 5 boroughs, Manhattan has the largest
population. Manhattan has more than 2.5 million residents, making up about 30% of
NYC population.

Due to its high population density, residents in NYC have low private car
ownership. The overall car ownership in NYC is 45% and only 22% in Manhattan. In
addition, only 8% of Manhattan residents drive to work. Taxi, subways, buses, and recently ridesharing, play essential roles in New Yorkers’ daily mobility.

Figure 2.5 Travel demand by taxi in NYC of January 2009

This study uses the taxi trip records from 2009 in NYC area provided by the New York City Taxi & Limousine Commission (NYC TLC), the major taxi company operating the famous yellow taxicabs in NYC. Based on the data from NYC TLC, a total of more than 143 million taxi trips were completed in 2009. For each taxi trip, location (latitude and longitude) and time for pick-up and drop-off were recorded, but the route was not included. Since we only use the pick-up and drop-off locations in this study, all
other trip information was not included in the analysis and discussion. Figure 2.5 shows the travel demand by taxi in NYC on January 1st, 2009.

We divided the study area into a cell size of 110-meter (0.001-degree latitude) by 80-meter (0.001-degree of longitude). After this step, NYC is divided into 87,838 cells and each cell is assigned with a unique identification number. Then, for each day, we count the demands of traveling for each cell as the summary of taxi trips starting from and ending in the given cell. We consider both pick-ups and drop-offs as demands for traveling and therefore both are treated the same. For example, for a given cell in a given day, there were 100 pick-ups and 200 drop-offs, which we consider 300 demands of traveling. After this step, all taxi travel data are organized as a matrix $A$. The entry $a_{ij}$ means the demand of traveling for the $i$th cell on the $j$th day.

2.5 RESULTS AND DISCUSSION

To validate the performance of this method, we apply this method to 2009 and 2012 NYC taxi data. For the 2009 model, we train our model on 2009 data and validate this model using a combination of 2010-2012 data. For the 2012 model, the training dataset is 2012 data, and validation dataset is a combination of 2009-2011. We first generate models with a different number of sensors and plot the overall $RMSE$ for each model. Figure 6 shows the $RMSE$ plotted to the corresponding number of sensors. For the training datasets, the general trend is that $RMSE$ keeps decreasing as the number of sensors increases. However, for validation datasets, the $RMSE$ reaches a lowest point. This lowest point is the optimal number of sensors we use in the model. Figure 6 shows the training and validation $RMSE$s for 2009 (Figure 2.6(a)) and 2012 (Figure 2.6(b)).
Based on these two figures, the optimal number of sensors for 2009 was 13 and for 2012 was 8.

Figure 2.6 Determination of the optimal number of sensors for 2009 (a) and 2012 (b)

2.5.1 Spatiotemporal events in 2009

Table 2.1 and Figure 2.7 illustrate the locations of the 13 sensors for 2009. The most important is Penn Station, which is the main transportation hub in Manhattan. Other sensors’ locations include some important intersections in Manhattan, multiple locations near the LGA airport, and other landmarks in Manhattan. Noticeably, other than 3 sensors located near the LGA airport, all other 10 sensors are in Manhattan, indicating that Manhattan travel demands are the dominant patterns to analyze the whole of NYC.
Table **Error! No text of specified style in document.** Locations of the 13 sensors for 2009

<table>
<thead>
<tr>
<th></th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Penn Station</td>
</tr>
<tr>
<td>2</td>
<td>Intersection of Essex St. &amp; Rivington St.</td>
</tr>
<tr>
<td>3</td>
<td>1st Ave outside Tisch Hospital/NYU Langone Hospital</td>
</tr>
<tr>
<td>4</td>
<td>Intersection of W 42nd St. &amp; 8th Ave. near Port Authority Bus Terminal</td>
</tr>
<tr>
<td>5</td>
<td>Javits Center</td>
</tr>
<tr>
<td>6</td>
<td>LGA dropoff area</td>
</tr>
<tr>
<td>7</td>
<td>LGA dropoff area</td>
</tr>
<tr>
<td>8</td>
<td>Lincoln Center for the Performing Arts</td>
</tr>
<tr>
<td>9</td>
<td>LGA taxi pickup area</td>
</tr>
<tr>
<td>10</td>
<td>Madison Square Garden</td>
</tr>
<tr>
<td>11</td>
<td>9th Ave between W 13th St and W 12th St</td>
</tr>
<tr>
<td>12</td>
<td>Empire State Building</td>
</tr>
<tr>
<td>13</td>
<td>9th Ave &amp; W 14th St.</td>
</tr>
</tbody>
</table>

Figure 2.7 Locations of 13 sensors in NYC for 2009 event detection

We then take the simulated scenario $\tilde{A}$ using 13 sensors to calculate daily $RMSE$, which can identify the daily anomaly over the whole year of 2009. The simulated matrix $\tilde{A}$ is organized in the same way as the original dataset $A$, in which each column represents one day’s travel demand of all cells. Therefore, calculating $RMSE$ column-wise is to compare daily difference
between the observed dataset and the simulated scenario. Result of daily \( RMSE \) is shown in figure 2.8.

![2009 Taxi Anomalies with 13 Sensors](image)

Figure 2.8 Daily \( RMSE \) showing the difference between observed travel demand vs simulated uneventful scenario for 2009

In 2009, the day has the largest difference between observation and simulated uneventful scenario was November 26\(^{th}\), which is Thanksgiving Day. Following by June 28\(^{th}\), the Pride March, November 1\(^{st}\), the day for NYC marathon, and December 25\(^{th}\), Christmas Day. As mentioned in Section 2.3.6, the Anomaly Index for each cell indicates the difference between our model simulation and the actual observation. By mapping this, we can find the spatial anomaly distribution.

Since most events in NYC are concentrated in Manhattan, as shown in Figure 2.9, in the results part, we only show the Manhattan map for a better illustration of anomaly distributions. We merged all the Anomaly Index numbers into one vector and applied the Jenks Natural Break classification (\textit{R: Jenks Natural Breaks Classification}, n.d.) to determine the ranges of each
category. In this way, the same color in all the maps represents the same range, which is more convenient for comparison across different time periods.

Figure 2.9 Spatial distribution of Anomaly Index for Thanksgiving Day, 2009

Figure 2.9 shows the spatial distribution of the Anomaly Index for Nov. 26th, Thanksgiving Day of 2009. The two places with the largest positive Anomaly Index are the Metropolitan Museum of Art and American Museum of Natural History, with 1770 and 1500 more observed trips than estimated, respectively. Midtown around the Times Square experienced more travel demands on the Thanksgiving Day than estimated. However most other places in Manhattan experienced less travel demand than estimated. The place with the largest negative Anomaly Index is Port Authority, meaning less travel demands were observed than estimated.
Figure 2.10 Anomaly Index of Manhattan area in 2009 for (a) June 28th (Pride March celebrating the LGBTQ community), (b) Nov. 1st (NYC Marathon), and (c) Dec. 25th (Christmas).

The day with the second largest overall \textit{RMSE} was June 28\textsuperscript{th}, 2009, which was the day for the NYC Pride March celebrating the LGBTQ community. Anomaly Index distribution is shown in Figure 2.10(a). 2009 was the 40\textsuperscript{th} anniversary of the Stonewall Riots in NYC. Although the Metropolitan Museum of Art and American Museum of Natural History have positive Anomaly Index, most of the cells with positive Anomaly Index are clustered around Midtown Manhattan near Times Square. Most areas in the Upper East Side and the Upper West Side are in blue, meaning observed travel demands are less than estimated.

Figure 2.10(b) shows the third largest \textit{RMSE} which appeared on Nov. 1\textsuperscript{st} when the NYC Marathon took place. This was the 40\textsuperscript{th} annual marathon race in NYC. The end of this race was southeast of Central Park. Therefore, the travel demand was less than estimated due to road closure for the race. Also, the last part of marathon route followed along the east side of Central Park. The blue area on the east part of the Central Park indicates the reduction of taxi travel.
demand due to the road closure. Cells with the largest Anomaly Index are along the 2nd Avenue in the Upper East Side and areas at the south tip of Manhattan, near the Battery waterfront park, where ferry to Statue of Liberty departures. This may be caused by tourism attracted to NYC for the Marathon events.

The fourth largest RMSE appeared on Dec. 25\textsuperscript{th}, which is Christmas day (Figure 2.10(c) shows the Anomaly Index). The majority of Midtown Manhattan has a negative Anomaly Index, meaning the number of actual trips is smaller than estimated. Midtown is the center of commerce. The decrease of taxi trips was caused by Christmas when most of companies in Midtown were closed. Both the Upper East Side and the Upper West Side had more trips than estimated on Christmas. There are two cells with an Anomaly Index larger than 1000, meaning observed travel demands were more than estimated. One cell is located on the 2nd Ave., between E 31st St. and E 32nd St, and the other one is located on Broadway between W 67th St. and 68th St. Both cells are near some multi-floor residential condominium buildings. Such higher travel demands were likely caused by Christmas visitations to or by nearby residents. The Metropolitan Museum of Art and American Museum of Natural History had a negative Anomaly Index. They were closed on the Christmas day and thus, less travel demands were observed.

2.5.2 Spatiotemporal events in 2012

Using the anomaly detection method for 2012, we found the optimal number of sensors was 8, when the validation dataset has the smallest RMSE (Figure 2.6 right). Figure 2.11 and Table 2.2 show the locations of these 8 sensors for 2012. Similar as observed in 2009, Penn Station was the most important sensor location, meaning it indicates the most correlations among taxi travel demands. Similar to 2009, the LGA taxi pick up area was also identified as a key location to indicate NYC taxi travel demands. In addition, the JFK airport terminal 4 departure
area was found to be a key location. Other than these two airports, all other key locations are all in Manhattan.

Table **Error! No text of specified style in document.** 2 Locations of the 8 sensors for 2012

<table>
<thead>
<tr>
<th></th>
<th>Locations of the 8 sensors for 2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Penn Station</td>
</tr>
<tr>
<td>2</td>
<td>Park Ave &amp; E 53rd St</td>
</tr>
<tr>
<td>3</td>
<td>9th Ave &amp; W 16th St - Google building</td>
</tr>
<tr>
<td>4</td>
<td>JFK Terminal 4 departure</td>
</tr>
<tr>
<td>5</td>
<td>Broadway &amp; W 66th St</td>
</tr>
<tr>
<td>6</td>
<td>11th Ave &amp; W 39th St - Lincoln Tunnel Manhattan end</td>
</tr>
<tr>
<td>7</td>
<td>LGA taxi pickup area</td>
</tr>
<tr>
<td>8</td>
<td>2nd Ave &amp; E 52nd St.</td>
</tr>
</tbody>
</table>

**Figure 2.11 Locations of the 8 sensors for 2012**
Figure 2.12 Daily RMSE showing the difference between observed travel demand vs simulated uneventful scenario for 2012

Figure 2.12 shows the daily RMSE during 2012. The day with the largest RMSE was Thanksgiving Day (Nov. 23rd). The second largest RMSE happened on the day when Hurricane Sandy hit NYC (Oct. 29th). The third largest RMSE appeared on the weekend for St. Patrick’s Day (March 18th). The day with fourth largest RMSE was New Year’s Eve (Dec. 31st).

Thanksgiving Day in 2012 also experienced a high overall RMSE, which was generated by several locations with extremely more travel demand than estimated. Figure 2.13(a) shows the Anomaly Index distribution for Thanksgiving Day. Locations with more travel demand than estimated were gathered in Midtown Manhattan, around the Fifth Avenue and Times Square. Most of Downtown Manhattan appeared in blue, meaning less travel demand were observed than estimated. Figure 2.13(b) shows the Anomaly Index distribution for the day when Hurricane Sandy hit NYC (Oct. 29th). Most of Midtown Manhattan appeared in blue. Midtown is where most companies were located. Many companies were closed and allowed work-from-home on that day. This mean that travel demand for Midtown was less than estimated. The Upper East
Side, which is mostly a residential area, had more travel demand than estimated. Higher travel demand was also seen near the NYU Langone Health Hospital and other residential areas. As many subways were closed due to potential flooding, more people chose taxi as the substitute travel mode.

Figure Error! No text of specified style in document. 2 Anomaly Index of Manhattan area in 2012 for (a) Nov. 23rd (Thanksgiving Day), (b) Oct. 29th (Hurricane Sandy) (c) Mar. 18th (St. Patrick’s Day), and (d) Dec. 31st, 2012 (New Year’s Eve)

Figure 2.13(c) shows the Anomaly Index distribution for St. Patrick’s Day. Fifth Avenue was overestimated and Madison Avenue, the next avenue parallel to the Fifth Avenue was underestimated. This happened because the Fifth Avenue was closed for the parade and thus taxi pick-up or drop-off activities were happened around the parade route. More travel demand was also observed near Times Square and surrounding commercial areas. Figure 2.13(d) shows the Anomaly Index distribution for New Year’s Eve. Areas near Times Square were in blue, meaning less than estimated travel demands were observed as the area was closed for traffic during the celebration events. Areas surround the Times Square had more than estimated travel
demand, as people who traveled to/from Times Square had to start or end their trips there. In addition, more travel demands were found near the Downtown Financial District and Battery Park, where the ferry to the Statue of Liberty leaves.

2.6 CONCLUSION

This chapter presents a new method based on Discrete Empirical Interpolation Method (DEIM) for event or anomaly detection using point-based human mobility data. This method can first identify the important locations in the study area and then simulate an uneventful scenario based only on limited observation data from the previously identified key locations. Spatiotemporal events or anomalies are detected by comparing the discrepancy between the observed actual data and the simulated uneventful scenario. Since this method is based on an unsupervised method, it does not require any prior knowledge of the study area or specification of an interested time window. In addition, this method requires less data preparation. Because the simulation process is based on each discrete temporal unit, it does not require chronologically ordered data, which can reduce pre-processing time.

This novel anomaly detection method described in this section demonstrates the feasibility of using optimal sensors-based simulation method for spatiotemporal anomaly detection. In this paper, locations of sensors are determined by principal components. Since the simulation process is based on observation data at those sensors’ locations, events taking place at the sensors’ locations cannot be detected. Future work can explore different methods to determine the sensors’ locations. In addition, future work can further utilize this method with other data types, such as social media data or cell phone data. In this work, spatial units and temporal units are arbitrarily determined. Future work can further explore different spatial and temporal unit combinations.
CHAPTER 3

SPATIOTEMPORAL PATTERNS OF HUMAN MOBILITY AND THEIR ASSOCIATION WITH LAND USE TYPES DURING COVID-19 IN NEW YORK CITY²

3.1 INTRODUCTION

The coronavirus disease 2019 (COVID-19), caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), was originally detected in Wuhan, China, and rapidly spread worldwide. By June 11th, 2020, more than 7 million cases were confirmed from 216 countries, areas, or territories (WHO, n.d.). By the end of May 2020, 1,782,571 cases and 104,220 deaths were reported in the contiguous United States (CDC, n.d.). Due to its rapid spread and sudden onset of severe symptoms, the World Health Organization (WHO) has declared the COVID-19 outbreak as a Public Health Emergency of International Concern on January 30th, 2020, and a pandemic on March 11th, 2020. On March 12th, 2020, President Trump declared a National Emergency Concerning the COVID-19 outbreak (The White House, n.d.).

To combat the spread of COVID-19, unprecedented measures were taken all over the world. One of the most well-known non-pharmacological measures is to keep “social

distancing”, including staying at least six feet from other people and no mass gathering (CDC, 2020). Similar policies of constraining human mobility have been found effective in reducing COVID-19 transmission in China (Kraemer et al., 2020; Maier & Brockmann, 2020), South Korea (Shim et al., 2020), and Italy (Gatto et al., 2020). In the United States (U.S.), federal, state, and local governments imposed social distancing measures in March to slow down the spread of COVID-19. Since early March, governors of multiple states declared the State of Emergency as more cases were reported and announced public school closures, non-essential business closures, and cancellations of big events. All those social distancing measures have been found to reduce the COVID-19 growth rate (Courtemanche et al., 2020)

Within the U.S., New York City (NYC) was one of the hardest-hit places at the beginning of the pandemic. Since the first confirmed case on March 3rd, the number of cases grew rapidly. By June 12th, 2020, NYC had 210,538 confirmed cases, which was about one-tenth of the total reported cases in the U.S. (CDC, n.d.). Field hospitals were set up in multiple places, and the USNS Comfort hospital ship arrived in New York Harbor to increase hospital capacity. Governor Cuomo placed the “New York State on PAUSE” executive order that started to be effective on March 22nd (hereafter 3/22), which includes strict social distancing measures. However, before the executive order, public schools were closed, and the majority of activities in NYC were suspended.

In responding to the battle of COVID-19, multiple companies released data about human mobility to the public. For example, Google and Apple, two main companies owning millions of users’ location information, published their reports on human mobility during the pandemic. However, the mobility reports published by Google and Apple were
highly aggregated data and the raw data was not accessible to the public. On the other hand, social media platforms with open data policies and location-enabled posts were effective data sources for human mobility research (Pan et al., 2013b). This chapter aims to explore human mobility patterns during the COVID-19 pandemic in NYC with open social media data (Twitter). In addition, this chapter identified changes in human mobility patterns spatially and temporally for different land use types based on NYC open data. Specifically, this chapter addressed the following three questions:

1. What are the spatiotemporal human mobility patterns revealed by Twitter data at the tax lot level?

2. How do human mobility patterns change by different land use types?

3. How are the change patterns detected by Twitter different or similar to the change patterns detected by Google Community Mobility Report?

This chapter contributes to the efforts of battling COVID-19 by using big data to reveal human mobility patterns during this pandemic. In addition, this paper bridges the gaps of using open data to detect human mobility changes by different land use types and comparing human mobility patterns revealed by different data sources.

3.2. BACKGROUND

3.2.1 Social Media for Human Mobility in Epidemiological Studies

With the increasing prevalence of geo-enabled social media platforms, an increasing number of people have used social media, such as Facebook and Twitter, as communication tools and information sources (Nawaz et al., 2017). Due to its prevalence, social media data have been used for epidemiological studies. Existing studies have found that social media data are useful in identifying infectious disease outbreaks.
(Charles-Smith et al., 2015; L. Tang et al., 2018), analyzing sentimental reactions and responses (Lwin et al., 2020; B. Zhu et al., 2020), assessing risks (Liao & You, 2014), and understanding disease dynamics (Bansal et al., 2016; Ye et al., 2016). Before the COVID-19 pandemic, social media data have been used for multiple infectious diseases including Ebola (Fung et al., 2016; Lazard et al., 2015), Zika (Fu et al., 2016; Stefanidis et al., 2017), H1N1 (Chew & Eysenbach, 2010; Signorini et al., 2011), dengue (Albinati et al., 2017; Kraemer et al., 2018; Ramadona et al., 2019), and seasonal influenza (Allen et al., 2016; Broniatowski et al., 2013).

Geotagged social media data showed their potential in helping analyze and predict the spread of infectious diseases by deriving human mobility patterns from the data. For example, Albinati et al. generated a prediction model for dengue using Twitter data (Albinati et al., 2017). Kraemer et al. derived human mobility patterns from Twitter and analyzed spatiotemporal transmission variation of dengue in Lahore, Pakistan (Kraemer et al., 2018). Taking a different approach, Ramadona et al. developed a dynamic mobility-weighted incidence index to analyze the spread of dengue in Yogyakarta, Indonesia (Ramadona et al., 2019). Lai et al. found that mobility patterns retrieved from cellphones are valuable to assess health-related risks for travelers (Lai et al., 2020). Barlacchi et al. studied relationships between characteristics of human mobility patterns and whether or not influenza-like symptoms exist (Barlacchi et al., 2017). Souza proposed spatial scan statistic methods to identify infection risks from Twitter (Souza et al., 2019).

3.2.2 Big Data for Human Mobility in Response to COVID-19
During the COVID-19 pandemic, big data contributed to fighting against COVID-19 from multiple perspectives (Bragazzi et al., 2020; Franch-Pardo et al., 2020; Gasser et al., 2020; Z. Li et al., 2021; C. Yang et al., 2020; C. Zhou et al., 2020). In the early stage, studies have contributed to understanding the spatiotemporal distribution patterns and spread patterns of COVID-19 confirmed cases in different countries, including China (Guan et al., 2020; J. Huang et al., 2020), Iran (Ahmadi et al., 2020; Arab-Mazar et al., 2020), Italy (Giuliani et al., 2020), and South Korea (Shim et al., 2020). Due to the high infectious rate and long incubation period, movement tracking and contact tracing were among the key methods applied to constrain the spread of COVID-19 (Grantz et al., 2020; J. Huang et al., 2020).

Location-based contact tracing is usually achieved with mobile phone-based location information (Cho et al., 2020; Keeling et al., 2020). Studies have utilized mobile phone location data to detect human mobility changes in China (Y. Zhou et al., 2020), Spain (Cecilia et al., 2020), South Korea (Lee & Lee, 2020), the U.S. (S. Gao, Rao, Kang, Liang, & Kruse, 2020; S. Gao, Rao, Kang, Liang, Kruse, et al., 2020, 2020; X. Huang et al., 2021; X. Huang, Li, Lu, et al., 2020; Jeffrey et al., 2020; Kang et al., 2020; Kogan et al., 2021), and Brazil (Queiroz et al., 2020), to name a few. With more countries imposing “stay-at-home” and “keep-social-distancing” policies, big mobility data have been used to analyze the effectiveness of social and political efforts that focus on reducing human movement. For example, how different places responded to reducing-mobility-related orders in the U.S. were analyzed at the state level (X. Huang, Li, Jiang, Li, et al., 2020a) and county level (S. Gao, Rao, Kang, Liang, & Kruse, 2020) with mobile phone data. The effectiveness of interventions in different countries was examined with mobility data (Cotti
et al., 2020; S. Gao, Rao, Kang, Liang, Kruse, et al., 2020; Lai et al., 2020). At the individual level, changes in travel behavior have been detected, including transportation mode choice, trip destinations, trip distances, and durations (Ghader et al., 2020; J. Huang et al., 2020; J. Zhao et al., 2020).

Although social media data have been widely used for human mobility studies, limited studies regarding social media and human mobility under this COVID-19 pandemic are found (Bisanzio et al., 2020; X. Huang et al., 2020; X. Huang, Li, Jiang, Li, et al., 2020a; Porcher & Renault, 2020). Existing COVID-19 and social media studies focus more on information propagation (Bridgman et al., 2020; Cinelli et al., 2020; Depoux et al., 2020; Kouzy et al., 2020; L. Li et al., 2020) and text-based analysis of specific topics (J. Gao, Zheng, Jia, Chen, Mao, et al., 2020; Koh & Liew, 2020; D. Li et al., 2020; Wahbeh et al., 2020).

3.3 STUDY AREA AND DATA

3.3.1 New York City

NYC includes five boroughs, with each borough being a county in New York State: Bronx (Bronx County), Brooklyn (Kings County), Manhattan (New York County), Queens (Queens County), and Staten Island (Richmond County).

The NYC Department of City Planning published extensive land use and related geographic feature data at the parcel level called PLUTO (Primary Land Use Tax Lot Output) (New York City Department of City Planning, 2020). PLUTO data include 857,205 parcels in total.

In PLUTO, all the parcels of NYC are categorized into one of these 11 land use types: one- and two-family buildings, multi-family walk-up buildings, multi-family
elevator buildings, mixed residential and commercial buildings, commercial and office buildings, industrial and manufacturing, transportation and utility, public facilities and institutions, open space and outdoor recreation, parking facilities, and vacant land.

3.3.2 Google Community Mobility Report

Google published a mobility report for COVID-19 by comparing changes in the daily number of visitors to the baseline. The baseline is composited of seven values, one for each day of the week, and the baseline value is the median value from a 5-week period (January 3rd to February 6th, 2020, hereafter 1/3 and 2/6). The Google Community Mobility Report includes changes for the following six categories of places: retail and recreation, grocery and pharmacy, parks, transit stations, workplaces, and residential. Changes reported in the Google Community Mobility Report are calculated as change percentages compared to the baseline. For the United States, the Google Community Mobility Report is available at the state- and county-level, if enough data are available (Google LLC, n.d.).

3.3.3 Twitter Data

Geotagged tweets were collected from 1/3, to May 30th, 2020 (hereafter 5/30), using the Twitter Stream Application Programming Interface (API). Although Twitter API only allows access to about 1% of all publicly available tweets (Pfeffer et al., 2018), a total number of 9,459,952 geotagged tweets from 209,775 users were collected from NYC during this 5-month period. All the streamed tweets are stored on a high-performance computing cluster. Queries for this study were conducted with Apache Impala and Apache Hive. Following the change calculation method in the Google Community Mobility Report, tweets collected between 1/3 and 2/6 were used for baseline calculation. Later tweets were used for mobility change calculation.
3.4 METHODOLOGY

3.4.1 Twitter Data Processing

We collected a total of 1,750,385 tweets between 1/3 and 2/6, 2020, for baseline calculation, and a total of 5,335,313 tweets were collected between February 16th (hereafter 2/16) and 5/30 for human mobility pattern examination. For all tweets we collected, we first separated them into daily datasets based on posted time so that one dataset only contains tweets posted during one specific day. In the second step, we mapped each tweet based on the associated geotag as a point. If the associated geotag was latitude-longitude coordinates, this tweet was directly located to the coordinates. If the geotag type is a place, for example, Central Park, this tweet is located to the centroid of central park. Tweets that cannot be mapped into tax-lot polygons were eliminated in this step. After this mapping step, 1,649,520 tweets remained for baseline calculation, and 5,162,468 tweets remain for human mobility pattern examination. In the next step, we count how many users have tweeted from each tax lot for each day. In other words, only one user is counted into the given lot, regardless of the number of tweets one user posted from the same location.

3.4.2 Spatial Patterns of Human Mobility Changes

PLUTO data for NYC had delineated each polygon for each tax unit with one of the 11 land use types listed in Section 3.1. Spatial mobility pattern changes are calculated for each tax unit.

The baseline used for change calculation is a 5-week period, from 1/3 to 2/6, which is the same as the Google Community Mobility Report. For each land unit, the
total number of Twitter users per week was summarized. The baseline value for each land
unit is the median of the five weeks’ weekly user amount.

For each polygon, the weekly Twitter user number was summarized for each week
between 2/16 and 5/30. The percentage of human mobility change of a given polygon was
calculated as Eq 3.1:

\[
change \text{ percentage} = \begin{cases} 
\frac{\text{observed value} - \text{baseline}}{\text{baseline}} \times 100\%, & \text{if baseline} \neq 0 \\
100\%, & \text{if baseline} = 0 
\end{cases} 
\]

Eq. 3.1.

3.4.3 Daily Mobility Pattern Change by Land Use Type

The baseline for the daily Twitter-based human mobility change pattern was matrix
\( B \), pairing land use type and day of the week. An element \( b_{ij} \) in the baseline matrix \( B \) was
the baseline value for land use type \( i \) on \( j \)th day of the week. For each week of the 5-week
baseline time, a matrix \( W \) was constructed, where each element \( w_{ij} \) represents the number
of Twitter users in land type \( i \) on the \( j \)th day of this specific week. Therefore, we built 5
matrices for the 5-week baseline time, denoting as \( W_1, W_2, \ldots, W_5 \). The baseline value \( b_{ij} \)
was the element-wise median of \( W_1, W_2, \ldots, W_5 \).

Twitter-based human mobility change patterns were calculated on a daily basis
from 2/16 to 5/30. For each day between this time period, we first used \( n_{ij} \) as the observed
number of Twitter users in land type \( i \) on the \( j \)th day of that week. Then, the daily change
percentage was calculated as follows:

\[
change \text{ percentage} = \frac{n_{ij} - b_{ij}}{b_{ij}} \times 100\%. \quad \text{Eq. 3.2}
\]

This change percentage calculation method can avoid fluctuations during the week,
such as human mobility pattern changes between weekdays and weekends.
Daily change patterns are presented as trend lines, where the y-axis is the change percentage, and the x-axis is the date.

3.5 RESULTS

3.5.1 Spatial Human Mobility Patterns based on Twitter Data

Figure 3.1 and Figure 3.2 show the human mobility patterns change detected using Twitter data at the parcel level. Figure 1 shows changes from week 1 to week 9, and Figure 3.2 shows changes from week 10 to week 15. From a spatial perspective, the decrease in the number of Twitter users can be observed as early as in the week of March 15\textsuperscript{th}, 2020 (hereafter 3/15, Week 5), which is the week before the city-wide executive order. From the Monday of week 5, most commercial and office buildings in Midtown Manhattan (dark blue polygons within the red rectangle in week 5) have seen more than a 75\% decrease in the number of Twitter users compared to January. Such reduction indicates the spontaneous actions of work-from-home from many companies. Downtown Manhattan (red circle in the figure of week 5), including the Soho District, Wall Street, World Trade Center, and the ferry to Statue of Liberty, has always been a popular destination of tourism. As the World Trade Center and Statue of Liberty suspended their tourism-related activities on 3/14 and 3/16, the number of Twitter users decreased more than 50\% at One World Trade Center and the Battery Park, which is where the ferry to the Statue of Liberty departs. After the stay-at-home executive order was placed on 3/22, the majority of non-residential buildings in Downtown Manhattan show more than a 75\% reduction in the number of Twitter users. In addition, Central Park (red rectangle in the figure of week 6) had seen more than a 50\% decrease in the number of Twitter users after the issue of the executive order, suggesting that the attraction of this largest park in New
York City has been dampened during the pandemic. The weekly number of Twitter users in Central Park remained at a similar level until the end of May. Contrasting to the general decreasing trend, hospitals (red circle in the figure of week 6) saw a significant increase in the number of Twitter users since week 6 (3/22), when the number of COVID-19 cases in NYC started to rise (CDC, n.d.).

Figure 3.3 Human mobility changing patterns at parcel-level in week 1 – week 9
Figure 3.4 Human mobility changing patterns at parcel-level in weeks 10–15.

3.5.2 Temporal Changes by Land Use Type

Figure 3.3 shows the daily change for the six land use types. The black line shows the original change pattern, and the red line is the smoothed trend with Gaussian smoothing.
Figure 3.5 Mobility changing patterns for six land use types based on Twitter data
Transportation-related lands show a significant decrease in the number of people—about a 70% decrease in mid-March. This result is expected, as people were encouraged to avoid unnecessary trips and stay at home. Commercial and office buildings were also found to have a noticeable decrease: about 60% in mid-March. Similarly, public facilities and institutions also observed about a 40–50% percent decrease in the number of Twitter users. Park and outdoor recreation areas for the whole of NYC indicated a slight decrease from mid-March. In April, about 40% fewer Twitter users were observed in park and outdoor recreation areas, compared to the baseline.

Twitter-based human mobility patterns show no obvious increase in the residential area for NYC. This shows one limitation of using Twitter data for this type of human mobility capturing study. We can only count the number of distinct Twitter users from a location. The length of time spent at home remains unknown. Therefore, for those users who regularly tweet from home before the pandemic, this number does not change.

The following sections describe human mobility changes for Bronx County, Kings County (Brooklyn Borough), Queens County, and New York County (Manhattan Borough) in detail. In addition, similarities and differences comparing human mobility change patterns from Twitter and Google are discussed. Richmond County (Staten Island Borough) is not further discussed, as it does not have enough data for a valid comparison.

3.5.2.1 Bronx County

Bronx County only has enough Twitter users for two land use types: residential and workplace. Figure 3.4 shows the comparison between Twitter-based human mobility pattern change and change generated with Google-based human mobility data.
In the Twitter-based human mobility pattern, the residential area shows no obvious increase, while Google shows a 20% increase in residential areas. As mentioned before, Google’s human mobility change for the residential area is based on time spent; therefore, a larger increase from Google is found. Both Twitter- and Google-based human mobility patterns find about a 60% decrease in the workplace in Bronx. However, the Twitter-based mobility pattern shows more dramatic day-to-day changes. Another major difference between these two trends is that Google shows notable weekday–weekend patterns.

Figure 3.6 Twitter- and Google-based human mobility changing patterns in Bronx County
3.5.2.2 Kings County (Brooklyn Borough)

Twitter-based human mobility patterns were generated and compared with Google-based human mobility change for Kings County in the following three land use types: residential, workplace, and park. Figure 3.5 shows the comparison for the three types.

We found a 20% increase in the number of Twitter users in Kings County for residential areas, which is similar to the Google Community Report. However, Twitter-based human mobility presents a larger day-to-day change, and this change is unstable. Google found that people spent about 20% more times in residential places, and weekly regularity was also found. Both Twitter- and Google-based human mobility patterns found about a 60% decrease in the number of people for the workplace in Kings County, and a drop starting around 3/15 was detected in both data.

However, Google also identified weekday–weekend regularity, which was not detected using Twitter data. The human mobility patterns for parks and outdoor recreational areas in Kings County are quite different between Google and Twitter. Google shows an obvious drop around 3/15, and the number of visitors started to increase in late April and continued increasing in May. The Twitter-based human mobility pattern shows an unstable pattern evidenced by the large day-to-day difference. In addition, the number of Twitter users identified in parks and outdoor recreation areas was above baseline for 71 days, even when the stay-at-home order was in place in late March and April.
Figure 3.7 Twitter- and Google-based human mobility changing patterns in Kings County
3.5.2.3 New York County (New York Borough)

New York County shows a decrease in the number of Twitter users in all categories (Figure 3.6). Noticeably, the total number of Twitter users in New York County has decreased by 30–40%. This was likely caused by the significantly decreasing tourism in NYC in general. The number of Twitter users for the residential area in New York County shows a 30% decrease, which is consistent with the number of Twitter users decreasing for the whole of New York County. Transportation-related lands show similar decreasing patterns. Both Twitter-based and Google-based human mobility patterns show a 70–80% decrease in visitors to transportation-related lands. The decrease of the number of Twitter users in transportation-related lands started in early March, which was earlier than the main drop that appeared in Google (mid-March). The workplace in New York County presents an obvious decrease in both Twitter-based and Google-based human mobility patterns. Twitter-based human mobility showed a 60% decrease, and Google showed a 70% decrease. One major difference is that the Google-based pattern shows obvious weekday–weekend regularity. Regarding parks and outdoor recreation lands, both Twitter-based and Google-based human mobility patterns identify the drop in the number of visitors around mid-March. However, the Google-based pattern shows a larger drop: about 70% until late April. Visitor numbers starting to increase in May were also found by the Google-based pattern. The Twitter-based human mobility pattern showed about a 40% decrease since mid-March. Unlike the Google-based pattern, no increase in May was found in the Twitter-based pattern.
Manhattan residential – Twitter

Manhattan residential – Google

Manhattan transportation – Twitter

Manhattan transportation – Google
Although Twitter- and Google-based human mobility patterns use different methods to calculate human mobility changes, both patterns showed a 20% increase for residential lands in Queens County (Figure 3.7). Similar to all residential lands, Google can identify weekday–weekend differences, while Twitter-based human mobility patterns cannot. Regarding transportation-related lands, both Twitter- and Google-based human mobility patterns show a significant drop in the number of visitors around 3/15. However,
the Twitter-based human mobility pattern shows a larger decrease: about 80%. On the other hand, the Google-based pattern found a 70% decrease, and the number of visitors started to increase in May, which was not identified by the Twitter-based human mobility pattern.

Figure 3.9 Twitter- and Google-based human mobility changing patterns in Queens County
3.6 DISCUSSION

3.6.1 Comparison of Mobility Patterns Derived from Google and Twitter

This chapter compares human mobility patterns during the COVID-19 pandemic using two different data sources: cell phone (Google) and Twitter. Human mobility patterns and changes of those patterns are different mainly because of the different data collection methods with these two data sources. The Google Community Mobility Report is generated using the Location History feature with Google Accounts. According to Google, if a user turns on Location History in their Google Account, and this user is signed into his/her Google Account on his/her mobile device, the location history is saved to Google. In other words, even if a user does not actively report location by posting or checking in with any Google products, Google is able to get this user’s location if this user has Location History turned on. With a large number of users who turned on Location History at the Google Account-level for the convenience provided by Google products, Google has access to continuous individual-level locations, which enables Google to analyze human mobility patterns at a more detailed and continuous level.

On the other hand, Twitter-based human mobility data collection relies on Twitter users’ activities completely. Individual-level location records are based on tweeting frequency and decisions about location sharing of each tweet for each individual user. Therefore, Twitter-based human mobility patterns can be viewed as discrete patterns. Comparing to human mobility data collected by Google, Twitter-based human mobility data fail to tell how long a user spends at a given location.

Mobility patterns in residential for Google Community Mobility Report and Twitter-based human mobility are calculated differently, which provides quite different
results. The Google Community Mobility Report calculates changes in the length of time spent in residential land, while Twitter-based human mobility calculates changes in the number of Twitter users. Therefore, the Google Community Mobility Report found increases in users on residential land for the four counties in NYC. Differently, Twitter-based human mobility indicates decreases in users on residential land in New York County (Manhattan). One reasonable guess is that residential Manhattan had a decrease in the number of visitors from outside of Manhattan, and those visitors contributed to a large portion of the tweets posted from the Manhattan residential area as they visit friends in Manhattan. Since the Google Community Mobility Report calculates changes for time length, longer time spent at residential land was found.

Workplace showed weekly cycles in the Google Community Mobility Report, but not in Twitter-based human mobility patterns. Interestingly, the Google Community Mobility Report showed more people visiting workplaces during the weekends. Such weekend peaks are found in Bronx County, Kings County, and New York County. However, Twitter-based workplace mobility patterns did not present such weekly cycles. Instead, the Twitter-based workplace mobility pattern in New York County showed fewer fluctuations after the lockdown.

In Queens County and New York County, human mobility patterns in transportation showed similar trends with Google- and Twitter-based human mobility patterns, although Google presents a smoother change than the Twitter-based human mobility pattern change. This was caused by the different number of sampled users. Since the Twitter-based human mobility pattern calculates changes with a smaller sample than the changes calculated in the Google Community Mobility Report, the resulting change
percentage is larger in Twitter-based human mobility. From this perspective, Google-based mobility patterns present changes in routine travel patterns such as commuting, while Twitter-based human mobility patterns are more sensitive to irregular patterns such as special events.

As demonstrated in this case, Twitter-based and cell phone-based data show different perspectives of human mobility patterns and changes of such patterns. Each data source shows specific perspectives of activity and human interaction with the land use types. Future research should consider more about data fusion to present multifaceted human mobility patterns and to capture a wider spectrum of the population.

3.6.2 Limitations

Although Twitter-based human mobility data provide fast and robust results for mobility pattern change detection, we realize that limitations exist in Twitter-based human mobility data. First is the issue of representativeness. Twitter, the same as all other social media platforms, can only attract a small portion of the total population. Furthermore, not all Twitter users post tweets with geotags to a resolution that can be valid for this study. Existing studies have recognized this problem and have examined for population biases with Twitter users (Hecht & Stephens, 2014; Y. Jiang, Li, & Ye, 2019b; Malik et al., 2015b). Unfortunately, no solutions have been found so far. Compared to Twitter users, Google has a larger number of users as sampled for mobility pattern changes. However, Google Location History requires additional action to be turned on. Similar to Twitter-based human mobility data collection, cell phone-based data are limited to users of certain companies, which is also affected by issues of representativeness (Wesolowski et al., 2013; Z. Zhao et al., 2016).
The second limitation of this chapter is the baseline calculation method. In this chapter, the baseline is calculated with a 5-week period data (1/3–2/6), which follows the Google Community Mobility Report. However, human mobility patterns during this time can be affected by seasonal tourism, especially considering NYC as an internationally famous tourism destination. For a better human mobility analysis, baseline calculation should consider the annual cycle of human mobility. For instance, human mobility data from last year March to May should also be included for baseline calculation.

3.6.3 Future Research

Despite the limitations, compared to most cell phone-based mobility data, the value of Twitter-based (or more generally social media-based) human mobility data lies in that they provide multi-dimensional information rather than only the changes in the location of individuals. Despite not being used in this study, tweets posted in NYC can be analyzed with text-based methods to retrieve information such as increasing situational awareness, allocation of medical needs, and reporting problems or change of services (Hooper et al., 2020; X. Huang et al., 2020). Such analysis can help us better understand human mobility dynamics during the pandemic, such as why people were still going to parks regardless of the stay-at-home executive order. In addition, with self-description and other information provided by each individual Twitter user, further analysis can be conducted, including how COVID-19 impacts different demographic and socioeconomic groups.

Secondly, Twitter data provide more flexibility in spatiotemporal resolution. Human mobility data provided by Google Mobility Report are at the county and state level. No finer resolution such as tract-level mobility patterns can be retrieved from those
Therefore, the spatial variance within a county cannot be captured. On the other hand, Twitter data can be aggregated to different spatial resolutions, which is a better fit for different research applications. Temporally, Twitter data allow individual-level mobility analysis, such as trajectory analysis and trip prediction. With long-term Twitter-based mobility data, more sophisticated change detection methods, such as time series, can be applied to better capture abnormality in human mobility.

3.7 CONCLUSION

This chapter explores the change of human mobility patterns in New York City using Twitter data. With the open access NYC detailed land use data, this study analyzes changes of Twitter user number of different land use types of NYC during the COVID-19 pandemic. This chapter also compares Twitter-based human mobility pattern changes with the Google Community Mobility Report at the county (borough) level of NYC. Comparing Twitter-based human mobility patterns and the Google Community Mobility Report, a major difference exists in residential lands, since Google Location History is able to record the length of time each individual spent at a location, which cannot be achieved with Twitter data. On the other hand, Twitter-based human mobility and the Google Community Mobility Report show similar results in changes at workplaces and transportation-related land use, such as subway stations and airports.

Human mobility data have played key roles in understanding and combating the COVID-19 pandemic. However, data accessibility placed limitations in scientific research, as many human mobility data sources are provided and protected by commercial companies whose raw data are not released to the general public. As an initial effort to bridge this gap, this study provides a comparison between the
effectiveness of identifying human mobility changes using social media data (Twitter) and cell phone data (Google) during the COVID-19 pandemic in NYC. The results of this study show that in some applications, open-access social media data (Twitter) can generate similar results to private data (Google), and thus, under some situations, social media data are capable of substituting private data to facilitate real-time human mobility monitoring. In addition, the results of this study can be used to develop platforms for infectious diseases monitoring and further analysis of different characteristics of human mobility patterns during the COVID-19 pandemic.
CHAPTER 4
SOCIAL DISTANCE INTEGRATED GRAVITY MODEL FOR
EVACUATION DESTINATION CHOICE

4.1 INTRODUCTION

Hurricanes are one of the most common yet costliest natural hazards in the United States. In 2016, hurricanes and associated heavy rainfall, storm surges, and strong winds caused death, injuries, and economic losses to coastal areas in the United States (National Oceanic and Atmospheric Administration (NOAA), n.d.). One of the primary mechanisms for protecting people from impending hurricanes and their hazards is evacuating the potentially affected area. Many disciplines, including geography, sociology, engineering, and political science, have contributed to a better understanding of evacuation behavior. Generally speaking, there are two different perspectives or foci for the research (Trainor et al., 2013). Transportation engineering studies focus on routing and destinations employing three-step models for evacuation: trip generation, trip distribution, and route assignment. The trip generation step models the evacuating population size and their response time (Herrera et al., 2019; Y. Zhu et al., 2018). The second step, trip distribution, models where trips end based on opportunities provided by each potential destination using origin-destination metrics (Bian et al., 2019; G. Cheng et

---

Then, in the last step, trips are assigned to different routes to optimize infrastructure usage (Bayram & Yaman, 2018; Ukkusuri et al., 2017).

Social scientists, however, focused more on individual- and household-level decision-making, aiming to gain a better understanding of how different factors affect evacuation decisions and behaviors. For example, studies have identified factors that affect evacuation decisions, including vehicle ownership, age, income, housing type, and other factors (Dow & Cutter, 1998, 2002; S.-K. Huang et al., 2016). Also, multiple social factors have been found essential to understanding evacuation time estimate (M. K. Lindell & Perry, 1992; M. Lindell & Prater, 2005), departure time distribution (S.-K. Huang et al., 2012, 2017), transportation mode (M. K. Lindell & Perry, 1992; M. Lindell & Prater, 2005), evacuation route choices (Dow & Cutter, 2002; Prater et al., 2000), and other evacuation logistics, including travel time, destinations, and accommodation (Bian et al., 2019; M. K. Lindell et al., 2011; H.-C. Wu et al., 2012). Findings of social factors in evacuation models have been reviewed by M. K. Lindell et al. (2018, 2020, 2020).

One challenge here is that some potentially useful social factors have not been examined in previous evacuation research. This is especially true for long-term related social factors. For example, how many counties has an individual visited in the last three years, and how much time he/she has visited each county? One potential solution is to collect social media data to retrieve digital footprints left by social media users. Social media data have been widely used in natural hazard-related research to understand evacuation behavior (Kumar & Ukkusuri, 2018; Martín et al., 2017b; Sadri et al., 2017). Besides the ability to provide rapid and easily accessible data, another advantage of social media is that long-term records can be retrieved. However, only limited studies have
utilized the long-term records from social media for evacuation behavior studies (Y. Jiang, Li, & Cutter, 2019).

This chapter aims to extend the functionality of social media data in evacuation behavior studies by utilizing users’ long-term traveling records from Twitter. Specifically, this chapter asks the following questions:

1) Are social media users more likely to evacuate to places they are familiar with?

2) How can social distance derived from social media data help to improve evacuation modeling?

To address these two questions, this chapter first reinforces findings from existing studies that social factors do play important roles in evacuation destination choices by quantifying the individual-level familiarity of each evacuated Twitter user, and then introduces a big data approach to measure county-to-county social distance based on geotagged tweets. Lastly, this study demonstrates how social distance can be integrated into gravity models to improve evacuation transportation planning.

4.2 LITERATURE REVIEW

4.2.1 Human Mobility Measured by Distance

The power law is one of the most commonly used distributions to model displacement distance in human movements. For example, the trip occurrence probability decreases as travel distance increases (Eq. 4.1), with the power law written as

\[ p(d) \sim \alpha d^{-\beta} \]

where \( p(d) \) is the trip occurrence probability, \( d \) is the trip distance, \( \alpha \) is a constant, and \( \beta \) is the scaling parameter (Brockmann et al., 2006; Mandelbrot & Mandelbrot, 1982).
Researchers further confirmed that the scaling parameter $\beta$ should be larger than 1 and smaller than 3. When $\beta > 1$, the trip occurrence probability forms an inverse proportional relationship with trip distance. When $\beta > 3$, this movement is Brownian, where the length of trip exhibits a Gaussian distribution (B. Jiang et al., 2009).

Benefitting from the prevalence of Global Positioning System (GPS) and location-enabled social media platforms, the availability of geotagged social media posts provides researchers with opportunities to advance understanding of human mobility. Noulas et al. (2011) collected 12 million user check-in data on Foursquare generated by more than 679,000 users in 111 days. This study provided an exploratory analysis of spatiotemporal distribution of users’ check-in locations for multiple categories of places. Similarly, Z. Cheng et al. (2011) collected about 22 million check-in data from nine social media platforms and modeled individuals’ travel distance with power-law distribution. Their power-law model agrees with previous human mobility studies using non-social media data sources (Brockmann et al., 2006; Gonzalez et al., 2008). In a cross-city study conducted by Noulas et al. (2012), more than 35 million trips were retrieved from Foursquare check-in data in 31 cities from different countries. This study shows that power law governs human mobility patterns in all the cities, though $\beta$ varies with city size and population density.

Human mobility patterns were also studied during previous natural hazards using social media data. Based on geotagged tweets, for example, Q. Wang & Taylor (2014) examined New York City residents’ daily travel patterns under the impact of Hurricane Sandy. This study found that although an individual’s activity center was shifted, their daily travel distances still follow a power-law distribution. The shift of the activity center
was caused by evacuation from flood-prone areas to safer areas. However, the shift of activity center was not the focus of their study and thus was not further examined. In another study by Q. Wang & Taylor (2016), they further confirmed their findings by testing whether human travel distance follows power-law distribution under multiple natural hazards. They collected Twitter users’ movement data during four typhoons (two in Japan and two in the Philippines), three earthquakes (in the Philippines, Chile, and the U.S.), three winter storms (in Britain, Germany, and the U.S.), three extreme rainstorms in the U.S., and two wildfires in Australia.

These studies provided two important contributions. First, although natural hazards caused perturbation for human movements, an individual’s travel distance distribution was still governed by the power-law distribution. Second, a shift of an individual’s activity space center can be observed, but the relationship between traveling center shift distance and an individual’s daily activity space was unrelated. These studies demonstrated the feasibility of using Twitter data to study human mobility patterns during natural hazards and the fitness of power-law distribution of travel distance during natural hazards. Also, the two studies by Q. Wang and Taylor (2014, 2016) revealed the shift in activity centers caused by hazard-related evacuations, but patterns of such shifts were not further examined. This latter point raised the question about whether evacuation distances of individuals still fit power-law distribution, which was examined in this paper.

4.2.2 Evacuation Destination Choices

Existing studies have developed multiple origin-destination models for evacuation transportation management at the aggregated geographical level. Evacuation destination choice is an important factor in determining evacuation transportation distribution.
Evacuation destination choice decisions vary among evacuees and are affected by multiple factors. Among those factors, accommodation is an important one that may decide evacuation destination. Common accommodation choices include, but are not limited to, friends’ relatives’ places, hotels/motels, and public shelters. Post-hurricane survey data showed that most evacuees chose to stay at a friend’s or relative’s home, as illustrated from evacuation behavior studies for Hurricane Floyd (G. Cheng et al., 2008) and Hurricane Ivan (Mesa-Arango et al., 2013). To examine factors that affect evacuees’ destination choices, G. Cheng et al. (2008) developed multinomial logit models for evacuees who went to friend’s/relative’s places and to hotels/motels. Their study found influential variables, including evacuation distance, whether the destination is affected by hurricane, population composition of destination, whether the destination is in a metropolitan statistical area, transportation convenience, and the probability of finding a place to stay at the destination. With evacuation-specified Traffic Analysis Zones (TAZs), Wilmot et al. (2006) developed three models for zonal aggregated evacuation destination choice. Comparing their gravity model, intervening opportunity model, and an extended intervening opportunity model that considered evacuation direction and hurricane path, only small differences were found (Wilmot et al., 2006). They further tested the transferability of the gravity model. The gravity model calibrated using data from Hurricane Floyd in South Carolina also worked with Hurricane Andrew in Louisiana (Wilmot et al., 2006). With the same survey data from Hurricane Floyd in South Carolina, G. Cheng et al. (2011) further extended the gravity model with dynamic features considering the storm path, road situation, and destination
accommodation availability. However, the transferability of this dynamic model has not been tested (G. Cheng et al., 2011).

Current gravity-derived trip distribution models calculate opportunities to each destination based on several factors, including pushing factors at origin, pulling at destination, and travel distance between origin and destination. Social factors included in existing models focus heavily on the pushing force at origins, such as information hurricane characteristics and evacuees’ vehicle ownership, risk perceptions of local residents (Dow & Cutter, 2002; M. K. Lindell et al., 2005), and pulling force at destinations, such as lodging options and cost (M. K. Lindell et al., 2011).

Since most evacuation studies rely on survey data, very limited long-term travel information can be collected. One of the advantages of using social media data is the availability of long-term data. For example, Y. Jiang, Li, & Cutter (2019) revealed that evacuated social media users have significantly larger long-term activity space than non-evacuated social media users.

4.3 DATA AND STUDY AREA

4.3.1 Hurricane Matthew and Its Affected Area

Hurricane Matthew was formed on September 28, 2016, and rapidly developed into a Category 5 storm, becoming the first Category 5 hurricane since 2007 in Atlantic basin (Stewart, 2017). It caused 585 direct deaths, with 34 in the United States. Based on the predicted track and the intensity of Hurricane Matthew, coastal residents from Georgia, South Carolina, and North Carolina were ordered to evacuate.

This chapter focuses on evacuation behavior of Twitter users living in these ten coastal counties before Hurricane Matthew: Chatham, GA; Brunswick, NC; Beaufort,
Berkeley, Charleston, Colleton, Dorchester, Georgetown, Horry, and Jasper in SC (Figure 4.1).

Figure 4.10 Hurricane Matthew path and the 10 selected counties

4.3.2 Data Collection and Preprocessing

Geotagged tweets were collected with the Twitter Stream Application Programming Interface (API) between July 2016 and December 2016. Streamed tweets were stored in a Hadoop environment and queried with Apache Impala in this study. We defined the resident county of a Twitter user as the county from which the user has posted the largest number of tweets (Y. Jiang, Li, & Cutter, 2019; Martín et al., 2017b, 2020; McNeill et al., 2017). From the massive Twitter dataset we collected, we identified local users whose resident county was one of these 10 counties.
The Twitter selection process followed Martín et al. (2017b) and Y. Jiang, Li, & Cutter (2019). Based on the predicted path of Hurricane Matthew, the governor of South Carolina issued a mandatory evacuation order on October 4th, 2016 (hereinafter “10/4”), followed by the governor of North Carolina and the governor of Georgia, who issued evacuation orders on October 6th. Given these evacuation orders, we considered the pre-evacuation time span as October 2nd, 2016 (hereinafter “10/2”) to 10/4, as evacuation was assumed to start after the evacuation order (10/4). The selected counties were under the impact of Hurricane Matthew between October 7th (hereinafter “10/7”) and October 9th (hereinafter “10/9”). We assumed the evacuation process had finished before the arrival of Hurricane Matthew (10/7) and that evacuees would not return before Hurricane Matthew had left (10/9). Therefore, we considered the post-evacuation time span as 10/7 to 10/9.

If a local user identified from the previous step posted during both pre-evacuation and post-evacuation periods, this user was collected for further analysis. Then, we compared each user’s posting location during the pre-evacuation and post-evacuation periods. If this user posted from within the 10 counties during pre-evacuation time and posted from outside the 10 counties during post-evacuation time, this user was considered an evacuated user. Also, each user’s pre-evacuation and post-evacuation locations must be at county level or finer to be considered a valid location. Users with state-level locations were eliminated as they could not be located in a specific county.

All the users we collected so far were manually checked to insure they were real, personally owned accounts. We used the Twitter API to retrieve the most recent 3,200 tweets those users had posted (3,200 was the maximum number of historical tweets
allowed to be queried by Twitter). Those who denied permission or deleted their accounts were eliminated from our dataset. Eventually, we collected 1,286 evacuated users with accessible historical tweets for further analysis. For each user, we defined the pre-evacuation location and post-evacuation as the centroid of this user’s geotagged tweets during pre-evacuation and post-evacuation time periods, respectively, and that evacuation distance is the driving distance from the user’s pre-evacuation location to post-evacuation location.

4.4 POWER-LAW DISTRIBUTION

Existing studies agree that power-law distribution governs individuals’ daily travel distance distribution during multiple natural hazards (Q. Wang & Taylor, 2014, 2016), but few tested whether evacuation distance follows a power-law distribution (Martín et al., 2017b). As evacuation distance is one of the most important factors for evacuation transportation planning, understanding the distribution of evacuation distance is an essential step.

The poweRlaw package (Gillespie, 2015) in R was used for this test. This package applies the bootstrap method to search for the best parameter using maximum likelihood estimation (MLE). The null hypothesis in this test was the evacuation distance distribution follows a power-law distribution. The bootstrap process converged at about 3,500 iterations and remained stable through 5,000 iterations. The estimated value was $\beta = 2.1776$ with a 95% confidence interval between 2.176 and 2.178. The scaling variable $\alpha = 115.03$. The final estimation generated a power-law distribution function as:

$$ f(d) = 115.03 \times d^{-2.1776} $$

Eq. 4.2
The p-value for the power-law fitness test was 0.526. As a result, we could not reject the null hypothesis that evacuation distance distribution follows a power-law distribution. Also, the resulting $\beta = 2.1776$ agreed with previous human mobility studies that $1 < \beta < 3$ (G. Cheng et al., 2011; Jurdak et al., 2015).

Figure 4.2 shows a histogram of evacuation distance distribution among the evacuees. There were a few evacuees who traveled less than 50 miles. They stopped immediately after they left the evacuation zone. Most of the evacuees evacuated to places about 150 miles away from coastal areas. That is about the distance from Charleston to Columbia in South Carolina. As one of the state evacuation strategies was to reverse lanes of Interstate Highway 26 (I-26) so that all the lanes of I-26 were directed from Charleston to Columbia to accelerate the evacuation process; 29.2% of evacuees from Charleston ended in the Columbia metropolitan area.

![Evacuation Distance Distribution](image)

Figure 4.11 Evacuation Destination Distribution
Unlike an ideal power-law distribution (the green line in Figure 4.2), the peak evacuation distance appears at the range between 100 and 150 miles, rather than at the beginning. Also, a slight bump can be found in Figure 4.2, about 600 miles at the distance. The power-law distribution of evacuation distances implicitly assumes that hotels, shelters, and other accommodations are uniformly distributed on a featureless space. However, in reality, accommodation opportunities are non-uniformly distributed. This explains why the evacuation distance distribution pattern differs from the power-law distribution.

4.5 FAMILIARITY MEASUREMENT USING TWITTER DATA FOR DESTINATION CHOICE

Survey data collected from multiple hurricane evacuations reported that over half of evacuees chose friends’ or relatives’ places as evacuation destinations (Bian et al., 2019; M. K. Lindell et al., 2018, p. 201; Mesa-Arango et al., 2013; Smith & McCarty, 2009). Mesa-Arango et al. (2013) developed a household-level nested logit model to analyze demographic and socioeconomic characteristics that affect evacuation destination choices based on survey data. Variables used in the model were directly from a survey, such as race, income, previous experience with hurricanes, and whether there was a need to work during evacuation (Mesa-Arango et al., 2013). Long-term travel behaviors were no discussed since they were not available from survey data. Bian et al. (2019) tackled this problem using community-level data from the American Community Survey (ACS) as a surrogate measurement for social factors. For example, length of living in the current community was used to measure the social network size in that study.
This section examines the relationship between evacuation destination choices and long-term social factors retrieved from social media. We used social media data to quantify the familiarity of destination for evacuees, using the assumption that people who evacuated to friends’ or relatives’ have a high degree of familiarity with that destination. Specifically, we focused on all the places an individual had visited before, and the likelihood that this individual would choose a place where he/she spent more time than a place he/she spent less time.

For all the evacuated users, we retrieved each user’s most recent 3,200 historical tweets; the maximum number of historical tweets allowed to access using Twitter API. An independent dataset was built to store each user’s historical tweets for further individual-level analysis. Then, we applied three steps to test these two hypotheses. First, for each evacuated Twitter user, we searched all the counties from which this user tweeted. Second, we retrieved all the available tweets for users identified from the previous step. As all the users’ historical tweets can be traced back to 2014, we identified how many days a user tweeted from each county since 2014 as tweeting frequency. Third, we ranked familiarity for all the counties from which this user had tweeted based on tweeting frequency identified in the previous step. For example, if an individual user tweeted from County A 200 days and County B 100 days, for this specific user, County A was ranked as the highest familiarity. If this user evacuated to County A, we counted this user as choosing the first in rank.

We excluded evacuation origin county from the familiarity rank, so this rank represents each user’s familiarity rank to evacuation destination county. In other words, the more days a user tweeted from the county, the more likely this user would evacuate to
the county. All the counties an individual had been to were ranked. This process was applied to all evacuated users, and their destination choice rank was summarized. Since the evacuation origin county, the residential county of each user, was excluded from the familiarity ranking, all the other counties included in the rank can be viewed as a potential evacuation destination for the specific user.

![Graph showing evacuation destination popularity vs familiarity rank.](image)

**Figure 4.12 Evacuation destination popularity vs familiarity rank**

Figure 4.3 shows evacuees’ destination choices. The x-axis is the familiarity rank, and each bar represents the percentage of evacuees who chose to evacuate to a county with a corresponding familiarity rank. Among 1,286 evacuees, 82.4% (1,060 evacuees) chose to evacuate to a county he/she had visited before. Specifically, 24.7% (318 evacuees) chose the county with the highest familiarity rank as evacuation destination. Also, 22.9% (295 evacuees) chose to evacuate to the county with the second highest familiarity rank. This result was further tested with Spearman’s rank order correlation test (Spearman, 1904). This test resulted in $p < 0.001$, indicating that the evacuee number and
familiarity rank are significantly negatively correlated, whereby the former decreases with the latter’s increase. For those who evacuated to a new place (marked as “no” in Figure 4.3), they may view this evacuation as an opportunity to explore a new tourism destination.

This analysis illustrated that familiarity with places is associated with evacuation destination decisions. Most evacuees chose their evacuation destination to be a county with a high familiarity rank. The higher familiarity rank a county has, the more likely an evacuation trip will occur.

4.6 IMPROVED GRAVITY MODEL

The gravity model is commonly used to model economic activities, trades, and human travel between a pair of places (Kepaptsoglou et al., 2010; Lewer & Van den Berg, 2008; Santana-Gallego et al., 2016). It can be written as Eq. 4.3:

\[ N_{ij} = G \frac{O_i^{\beta_1} D_j^{\beta_2}}{d_{ij}^{\alpha}} \]  

Eq. 4.3

When used for evacuation, \( N_{ij} \) represents the evacuation population from the origin \( i \) to the destination \( j \). \( O_i \) and \( D_j \) are the total population sizes of the origin \( i \) and the destination \( j \) respectively. The denominator part is a fringe function, interpreted as the cost from traveling between the origin \( i \) and the destination \( j \). In the traditional gravity model, the fringe function is based on physical distance \( (d_{ij}) \) between the origin \( i \) and the destination \( j \). As power-law distribution indicates, when the distance between two places increases, the probability of travel occurrence decreases. \( G \) is the gravitational constant, functioning as a scaling parameter. \( \beta_1, \beta_2 \) and \( \alpha \) are heuristic parameters for the origin population \( (O_i) \), the destination population \( (D_j) \), and the physical distance \( (d_{ij}) \).
Previous studies have examined the fitness of the gravity model and the intervening opportunity model. The relationship between the gravity model and the intervening opportunity model, and their extended forms, are reviewed by B. Chen (2005). Existing evacuation models only consider the physical distance \(d_{ij}\) as the difficulty of making the trip between each pair of origin and destination. As indicated in Eq. 4.3, an increase in the physical distance decreases the trip occurrence when other parameters are unchanged. We argued that social distance between a pair of places also functions in such gravity-based evacuation models. An increase in the social distance decreases trip occurrence when other parameters are unchanged. Section 4.6.2 provides a test of how social distance improves the accuracy of traditional gravity model. In this study, social distance was represented as the inverse of the familiarity measurement aggregated at county level, which was calculated as a social connection measurement.

4.6.1 Social Connection Measurement

Social connection measurements have been developed and used by multiple urban studies to measure the strength of connectivity between two places (Browning & Cagney, 2002; Zhong et al., 2014). Among the different variables used in the social connection model, human movements always play important roles, although different types of human movement data are deployed in different measurements.

The social connection measurement developed in this study was based on travels retrieved from Twitter users’ records. It represented the likelihood of a trip occurring between the given two counties in the long term. It was based on the assumption that the more the travels between two counties, the stronger the social connection between two counties, the shorter the social distance, and therefore, the more likely an evacuation trip
occurred. Specifically, the measurement was calculated as the percentage of Twitter users who traveled between the given two counties based on geotagged tweets collected in a six-month period (July 2016 to December 2016) following Eq. 4.4.

\[
Social\ Connection = \frac{N_{OD}}{N_O} \times 100\% ,
\]

where \(N_{OD}\) is the number of Twitter users found in both the origin county \(O\) and the destination county \(D\), and \(N_O\) is the total number of Twitter users in the origin county \(O\).

The calculation process involved four main steps. First, we identified users who sent tweets from the 10 coastal counties in the six-month period. Second, for each individual user, we found all the counties he/she had tweeted as a user’s active counties. If a user was active in more than one county, this user built a connection between each pair of active counties. For example, if a user posted geotagged tweets from County A, County B, and County C, connections were strengthened between Counties A and B, between Counties B and C, and between Counties C and A. In the third step, we aggregated to the county level. Since our focus was the social connection between the 10 counties to other counties, connections between a pair of counties within the 10 counties or a pair of counties not in the 10 counties were not calculated. Finally, the results from the previous step were divided by the total Twitter user of the origin county to standardize this measurement. For example, between July 1\(^{st}\) and December 31\(^{st}\), 2016, we found that 667 Twitter users had tweeted from both Brunswick County, NC, and Mecklenburg County, NC, and that the total number of Twitter users found in Brunswick County was 25,150. In this case, \(N_{OD} = 667\) and \(N_O = 25,150\). The social connection between Brunswick County (Wilmington in Figure 3.4a) and Mecklenburg County (Charlotte in Figure 3.4a) was 2.65\%, based on Eq. 4.4.
Figure 4.13 The social connection for the four selected counties

Figure 4.4 shows the social connections of (a) Brunswick County, NC, (b) Chatham County, GA, (c) Charleston County, SC, and (d) Horry County, SC. For better visual illustration, connections to some counties that are too weak to be visible or counties that are too far to be included in this map scale level were eliminated in this figure. The width of the red line represents the strength of social connection between two counties. In Figure 4.4a, Brunswick County has the strongest connection with Mecklenburg County, stronger than connections with other counties having shorter physical distances. Chatham County (Figure 4.4b) has the strongest social connection with Fulton County, GA. Although some other counties have shorter physical distance from Chatham County, social connection is actually weaker than the connection between
Fulton County and Chatham County. In Figure 4.4c, Charleston County has strong connections with counties near Columbia, SC. It also has a relatively strong connection with Fulton County, GA, and Orange County, FL. Both counties have larger physical distance than counties in South Carolina, but social connections with Charleston County are stronger. Figure 4.4d shows the strongest connection Horry County, SC, has with Mecklenburg County, NC. Also, it has relatively strong connections with counties near Nashville, TN. Figure 4.4 shows that social connections are not proportional to physical driving distance. Therefore, when modeling evacuation destination choice, the social connection should also be considered for inclusion in the fringe function to better model human mobility.

4.6.2 Social Distance Integrated Gravity Model

Social connection was integrated into the fringe function of the gravity model as an additional measurement of distance (considered as social distance) besides physical driving distance (Eq. 4.5). $f_{ij}^α$ represents the county-to-county social connection between the origin $i$ and destination $j$. $α_2$ is the heuristic parameter and $l$ is the scaling factor.

$$N_{ij} = G \frac{o_i^{α1}D_j^{α2}}{d_{ij}^{α1} + l f_{ij}^{α2}}.$$  Eq 4.5

Since the driving distance was used in this model as the physical distance, driving was the only transportation mode we considered in this study. Therefore, counties exceeding 1000 miles away from the 10 origin counties were eliminated. Also, counties without observed traveling with any of the 10 counties were also eliminated. Specifically, if a county did not receive any evacuees during Hurricane Matthew and no common users were found with any of the 10 counties in the 6-month period, this county was also eliminated even if it was within 1000 miles. This step eliminated 38 counties from the
observed 326 destination counties and left 288 counties for model calibration. The social connection was calculated using the method described in Section 4.6.1. A nonlinear optimization function was used in R to optimize scaling parameters \((G \text{ and } l)\) and heuristic parameters \((\beta_1, \beta_2, \alpha_1, \text{ and } \alpha_2)\).

For comparison, we first optimized the traditional gravity model (Eq. 3.3). The optimization was run in R with nonlinear model optimization. The optimized traditional model is as follows:

\[
N_{ij} = 8.477 \times 10^{-5} \frac{O_i^{0.61} D_j^{0.74}}{d_{ij}^{0.92}} \tag{Eq. 4.6}
\]

We conducted an exhaustive cross-validation of this model. This cross-validation process included two rounds of leave-one-out cross-validation (Molinaro et al., 2005). This process re-sampled all the data into training and testing datasets to avoid overfitting problems. We organized the dataset into the following table (Figure 4.5), where each column is an evacuation origin and each row is an evacuation destination.
Figure 4.14 Cross-validation process

The first round is leave-one-row-out cross-validation. This includes multiple runs. In each run, one row is left out as the test dataset. The remaining 287 rows are used to train the model (Eq. 4.3). After the model was finished training in each run, the one being left out was used to test model performance in this run. An RMSE was calculated by comparing the difference between the observed value and the output from the trained model. Since we have 288 evacuation destinations, the first round of cross-validation includes 288 runs and generates 288 RMSE values. The second round of cross validation is leave-one-column-out. In this round, we leave one column out as the test dataset and use the remaining nine columns to train the model (Eq. 4.6). Like the previous cross-validation round, an RMSE value is calculated in each round. The second round includes 10 runs, as we have 10 evacuation origins. Therefore, 10 RMSE values were generated in the second round of cross-validation. After two rounds of leave-one-out cross-validation,
a total of 298 RMSE values were generated. The overall average of RMSE for all the validation runs is 1.24, and the standard deviation is 1.68.

With the social distance integrated into the model, the improved gravity model is shown in Eq. 4.5. Like the previous model optimization process, the nonlinear model optimization procedure was run in R for the improved gravity model. The result is shown in Eq. 4.7.

\[
N_{ij} = 2.66 \times 10^{-5} \frac{O_i^{0.60} D_j^{0.70}}{d_{ij}^{0.90} - 0.950 D_{ij}^{0.83}} \quad \text{Eq. 4.7}
\]

Like the traditional gravity model, two rounds of leave-one-out cross-validation were conducted to avoid the overfitting problem. After two rounds of cross validation, a total of 298 RMSE values were generated. The overall average RMSE was 0.80 and the standard deviation was 1.88.

Comparing these two models, the improved gravity model reduced the overall average RMSE from 1.24 to 0.80, which was a 35% error reduction. In other words, the social distance integrated gravity model shows an improvement of 35% accuracy in predicting evacuation destinations comparing to the gravity model that only considered physical distance. This demonstrates the utility of social distance in evacuation destination prediction models and can be applied to practical applications, such as evacuation transportation planning.

4.7 LIMITATION AND FUTURE RESEARCH

Although the proposed model significantly reduced RMSE, we realized some limitations to this research. The first is the Twitter representativeness issue (Y. Jiang, Li, & Ye, 2019a; Malik et al., 2015a). Using Twitter data introduces population biases toward a certain group and may not represent all populations with various demographic
and socioeconomic characteristics. Although the representativeness issue of social media data is recognized and recent studies have advanced understanding of the demographic and socioeconomic characteristics of social media users using a different method, no unanimous solution has been reached. One potential solution is to develop a better sampling method that integrates both survey and social media data. For example, Martín et al. (2020) compared age, gender, and race between users collected from surveys and social media in evacuation studies. Integrating multiple data sources and developing a better sampling method are required for a better understanding of evacuation destination choices of different population groups.

The second limitation is variable choices. To demonstrate the functionality of social distance, the models in this study were only modified regarding the distance \(d_{ij}^\alpha\) in the gravity model (Eq. 4.5). Undoubtedly, distance draws the most attention in evacuation transportation planning, but it is not the only factor. Various other social factors identified in existing studies also play important roles in evacuation destination choices, such as family size, hotel/motel availability, financial budget, and more. These variables could be used to calibrate \(O_i^{\beta_1}\) and \(D_j^{\beta_2}\) in Eq. 4.5. The proposed model can potentially be further improved by integrating more variables in the optimization function.

The third limitation concerns the evacuation transportation mode. This study eliminated evacuees who traveled more than 1000 miles, a reasonable estimate of the maximum distance that households would travel by car. However, people with long distance travel during evacuation times were observed. Evacuees were observed to travel to the west coast, including Los Angeles and Seattle. How to integrate the multiple
transportation modes into the evacuation model optimization process requires further investigation.

4.8 CONCLUSION

This chapter responded to the calls for interdisciplinary models for evacuation behavior studies by improving current evacuation destination choice models through integrating social distance with traditional gravity models. It offered a potential solution to the challenge of lacking long-term data for essential social factors for evacuation behavior studies using a traditional data collection method (e.g., survey). The main contributions of this study came from the following three perspectives. First, this study reinforced and extended the important roles of social factors in evacuation modeling by confirming that familiarity with a previously visited place was associated with evacuation destination choice decisions. Second, it developed an approach to quantitatively measure county-to-county social distance using geotagged tweets. Third, it demonstrated how long-term social factors improved the evacuation destination choice model by integrating social distance into the gravity model.

Evacuation mobility patterns are complicated. Hardly could one generic mathematic model accurately represent such patterns. This study sheds light on how long-term traveling information retrieved from social media can quantitatively improve current transportation modeling for evacuation destination choice. With the increasing usage of social media during time-critical situations, methodological development in related research areas should be pushed further. Given the improvements observed in this study, we expect to see more studies using hazard-related social media data for evacuation model improvement.
CHAPTER 5

CONCLUSION

The study of human mobility has been challenging given the complexity of humans’ living environment. Although multiple studies have successfully found some universal patterns and can predict human mobility to some degrees, such patterns are based on uneventful regular movements (Gonzalez et al., 2008; Song, Qu, et al., 2010). In a complex living environment, such human mobility patterns may be disrupted by both expected and unexpected events.

The technological developments in smart phone and geolocation-related services have significantly changed the way researchers collect human mobility data. Various data collection methods have caught the attention of researchers, such as communications with cellular towers, Wi-Fi signals, geotagged social media posts, and in-vehicle GPS records (Toch et al., 2019; A. Wang et al., 2020). These new data collection methods overcome some limitations of traditional self-reported travel behavior data collection, especially during disruptive events. For example, protective evacuations for an upcoming natural hazard are mainly collected by post-disaster surveys, which are delayed, expensive, and labor-intensive (Martín et al., 2017b, 2020). Geotagged social media posts can record coastal residents’ evacuation behaviors during multiple hurricanes and thus serve as an innovative data source to understand evacuation behavior (Y. Jiang, Li, et al., 2021; Martín et al., 2017b; Mirbabaie et al., 2020; Sadri et al., 2018). In addition, small-scale observations may result in biased data, which could cause unreliable conclusions.
Large-scale human mobility data from in-vehicle GPS records could overcome this problem.

The availability of big human mobility data provides an unprecedented opportunity for researchers to conduct studies to advance understandings in human mobility. With more data collection methods, the volume of human mobility data kept growing, which requires advanced analysis and computing methods. The increase in human mobility data types balloons the data size and hastens the speed of new data generation. The fast rate at which new human mobility data are generated further increases the volume of data and brings more variety in human mobility data. While the “3Vs” of big data brought attention to data capturing and storing developments, methodological developments to understand the value of human mobility big data are lagging behind. New quantitative analytical methods are needed to mine the valuable information from human mobility data and to understand what human mobility data can tell us. Following this need, this dissertation contributes to the quantitative methodological development to understand human mobility patterns during disruptive events with geospatial big data.

The novel spatiotemporal event detection method proposed in Chapter 2 uses a data-driven quantitative method to first identify important locations in the study area, and then detect spatiotemporal events with a simulation method. This event detection method uses the Discrete Empirical Interpolation Method, which is a method for dimensionality reduction. Therefore, this method can handle a large volume of geospatial data at a fast rate. In addition, this event detection method simulates an uneventful human mobility pattern for each temporal unit based on very few observation points. Because of this, this
method does not require previous chronologically sorted data, which saves pre-processing time. In Chapter 2, this method is validated with New York City taxi travel records during 2012, but this spatiotemporal event detection method can be applied to other types of human mobility data (e.g., geotagged social media data) in different study areas and at different spatiotemporal scales.

Chapter 3 of this dissertation handles the variety perspective of big data by comparing two different data sources. Specifically, Chapter 3 compares an open social media data source – Twitter-based human mobility data – and a private human mobility data source with limited access – data released by the Google Community Mobility Report – during the COVID-19 pandemic. Results in Chapter 3 show that Twitter-based human mobility patterns are similar to Google-based human mobility patterns in most land use types. The major advantage of Google-based human mobility data is that Google is able to detect the length of time an individual stayed at a given location, which results in the difference in human mobility pattern changes found in residential areas. Chapter 3 emphasized the importance of open data sources, especially the timeliness of the near real-time social media data.

Chapter 4 of this dissertation tackles the research question on how to leverage historical social media data in understanding protective evacuation destination choice. Specifically, Chapter 4 improves the traditional gravity model for evacuation destination choice by integrating a social distance measurement. This social distance measurement is based on individual’s historical geotagged tweets and thus provides a useful measurement of how a user is familiar with a potential destination county. This improved gravity model was applied to the 2016 Hurricane Matthew evacuation as a case study. Results of
Chapter 4 show a 35% accuracy improvement in predicting county-to-county evacuation numbers. While many recent studies focused on how to quickly capture and query disaster-related information from near real-time social media data, Chapter 4 demonstrates the potential of long-term social media data in hurricane evacuation modeling.

Since Goodchild published “Citizens as sensors” in 2007 (Goodchild, 2007), volunteered geographical information (VGI) has been a hot topic in GIScience; recent geolocation services further advanced scientific methods in VGI collection. In the decade since, massive volumes and wide variety of geospatial big data are now being generated at a high velocity. GIScience researchers are equipped with skills to capture, query, and visualize such data. However, how to mine valuable information from geospatial big data still remains as a partially unanswered question. This dissertation is an effort to advance the study of quantitative methods to understand human mobility patterns with geospatial big data, but further research is required.

First, researchers need to better understand the representativeness geospatial big data. For example, Twitter is a widely used social media data source given the convenience of its open access, but it is a biased population sample. Multiple studies have identified the representativeness issue — that some demographic and socioeconomic groups are over-represented while some other groups are under-represented (Y. Jiang, Li, & Ye, 2019a; Malik et al., 2015a). When using social media to understand human mobility patterns, future studies are required for better sampling social media users or find complimentary data sources for less biased population samples. In addition, biases in other data sources also need to be identified. For instance, users may have different
preference in cellphone carriers or navigation applications. These potential biases existed in big data may impact the resulted human mobility patterns.

The second direction for future research is in data comparison and data fusion. The third chapter of this dissertation compares multiple data sources in order to understand the advantages of different data sources. Future research can conduct statistical tests to further validate these comparison results. In addition, future quantitative methodological developments can focus on methods that are able to take multiple data sources as inputs and leverage the advantages of each data source to generate optimal results. Future research should also focus on cross-validation of different data sources for human mobility to check whether or not different data sources can find the same patterns.

Other future improvements include optimizing parameters used in this dissertation. For example, in Chapter 2, different spatial and temporal units may result in different event detection results. How to better determine these units requires more sophisticated consideration. This also involves the classic Modifiable Area Unit Problem (MAUP). Although some studies have tried to tackle this problem (Fotheringham & Wong, 1991; Ye & Rogerson, 2022), more adjustments and analysis are needed.

This dissertation explores the potential of geospatial big data to understand human mobility patterns during disruptive events by developing a set of quantitative methods. Current research using geospatial big data, especially publicly accessible geospatial big data, in human mobility is still at an exploratory stage. Previous studies focused on collecting, processing and visualizing human mobility related big data, but only limited studies have tailored quantitative methods specifically for human mobility applications.
Quantitative methodological contributions from this dissertation aim to advance scientific research in terms of understanding how to mine information from big human mobility data. Thus, such information can be used for practical applications, including but not limited to emergency management and urban planning.
REFERENCES


Barbosa, H., Barthelemy, M., Ghoshal, G., James, C. R., Lenormand, M., Louail, T.,

getting sick? Predicting influenza-like symptoms using human mobility
behaviors. *EPJ Data Science, 6*(1), 27.

Barnes, T. J. (2004). Placing ideas: Genius loci, heterotopia and geography’s quantitative

https://doi.org/10.1191/0309132504ph506oa


https://doi.org/10.1126/science.1151419

under uncertainty: A benders decomposition approach. *Transportation Science,
52*(2), 416–436.

hurricane evacuation mode and destination type joint choice using data from
multiple post-storm behavioral surveys. *Transportation Research Part C:
Emerging Technologies, 99*, 130–143.

Bisanzio, D., Kraemer, M. U., Bogoch, I. I., Brewer, T., Brownstein, J. S., & Reithinger,
R. (2020). Use of Twitter social media activity as a proxy for human mobility to
predict the spatiotemporal spread of COVID-19 at global scale. *Geospatial
Health, 15*(1).


https://doi.org/10.1111/j.1541-0064.1963.tb00796.x

https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-
Deaths-by-State-o/9mfq-cb36

getting-sick/social-distancing.html

Mobile crowdsensing approaches to address the COVID-19 pandemic in Spain.
IET Smart Cities, 2(2), 58–63.

Charles-Smith, L. E., Reynolds, T. L., Cameron, M. A., Conway, M., Lau, E. H., Olsen,
social media for actionable disease surveillance and outbreak management: A
systematic literature review. PloS One, 10(10), e0139701.

2012 IEEE 12th International Conference on Data Mining, 141–150.

intervening opportunity model.

small data for travel behavior (aka human mobility) analysis. Transportation

Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. Mobile Networks and


Jeffrey, B., Walters, C. E., Ainslie, K. E., Eales, O., Ciavarella, C., Bhatia, S., Hayes, S., Baguelin, M., Boonyasiri, A., & Brazeau, N. F. (2020). Anonymised and aggregated crowd level mobility data from mobile phones suggests that initial compliance with COVID-19 social distancing interventions was high and geographically consistent across the UK. *Wellcome Open Research, 5*.


http://www.scopus.com/inward/record.url?scp=0019707252&partnerID=8YFLogxK


ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, 344–353.


