ABSTRACT

Title of Dissertation:	Introducing Frameworks to Analyze Human Mobility Behavior with Advanced Computational Algorithms and Machine Learning Methods Using Mobile Device Location Data
	Aref Darzi, Doctor of Philosophy, 2022
Dissertation directed by:	Professor Lei Zhang, Department of Civil and Environmental Engineering

The emergence of mobile device location data (MDLD) provides new opportunities to analyze human mobility behaviors. The large penetration rate and the possibility of observing human mobility behaviors continuously are among the most important features of the passively collected mobile device location data. However, to utilize MDLD in mobility behavior analysis, comprehensive computational algorithms need to be developed to carefully process the data.

This research proposes novel sets of frameworks to extract mobility context from the raw MDLD. First, this study introduces a set of algorithms to construct the travel

behavior of mobile device owners along with the non-observable attributes of both trips and travelers by extracting trips, identifying significant activity locations of the travelers such as their home and work locations, and imputing the travel mode. The proposed algorithms in this study were tested against the state-of-practice and state-of-art algorithms developed in the literature. The proposed algorithms were shown to have superior performance compared to other methods.

Next, this study further examines the usefulness of the proposed framework in providing near real-time insights on the evolution of human mobility behavior during the Coronavirus disease 2019 (COVID-19) pandemic. As a part of this study, a new metric has also been introduced to measure the social distancing practices from the mobility perspective. Additional investigations are also conducted to understand the linkage between the outbreak of COVID-19 and the mobility behavior of the communities.

Lastly, this study seeks to develop a framework to investigate the evacuation behavior of individuals during a natural disaster and construct the evacuation evolution patterns and decisions based on the MDLD. This dissertation evaluates the importance of the historical mobility behavior of the device owners in their decision-making procedure during natural disasters using statistical discrete choice models.

INTRODUCING FRAMEWORKS TO ANALYZE HUMAN MOBILITY BEHAVIOR WITH ADVANCED COMPUTATIONAL ALGORITHMS AND MACHINE LEARNING METHODS USING MOBILE DEVICE LOCATION DATA

by

Aref Darzi

Dissertation submitted to the Faculty of the Graduate School of the

University of Maryland, College Park, in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy
2022

Advisory Committee:
Professor Lei Zhang, Chair
Professor Katharine Abraham, Dean's Representative
Professor John C. Haltiwanger
Professor Deb A. Niemeier
Professor Erkut Y. Ozbay

Acknowledgments

I am so grateful to all who have supported me throughout this journey. I would like to express my deepest appreciation to my advisor and committee chair, Professor Lei Zhang, for his continuous support, vast knowledge, and inspiring advices throughout my Ph.D. study at the University of Maryland. Without his guidance and persistent help, this dissertation would not have been possible.

I would also like to thank my committee members, Professor Katharine Abraham, Professor John C. Haltiwanger, Professor Deb A. Niemeier, and Professor Erkut Y. Ozbay for their encouragement, insightful comments, and valuable suggestions to my research. I have learned a lot during the past several years from all of you and I would not be where I am without you all. Thank you.

This dissertation work would not have been possible without the help and guidance of my colleagues, Dr. Sepehr Ghader, Dr. Chenfeng Xiong, Dr. Yixuan pan, Mofeng Yang, Qianqian Sun, Aliakbar Kabiri, and Guangchen Zhao. I am truly grateful to them, and all of those with whom I have had the pleasure to work during the past several years at the University of Maryland. I would like to also use this opportunity to thank all my friends who have been like family to me, especially Shahrzad Saffari who has been there to support me since the first day of my Ph.D. journey.

I would also like to thank Maryland Transportation Institute at the University of Maryland for their financial support.

I could not have done this without the unconditional love and support of my family. Special thanks to my parents Ali and Khadijeh, and my siblings Hadi, Hamed, and Maedeh. Thank you for always being there for me and showing me what is truly important in life.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1.Overview

Understanding the mobility pattern of humans both at individual and aggregated levels is an integral part of transportation studies (1, 2). Analyzing people's mobility behavior provides important inputs such as how, from where to where, and when people travel for planners and decision-makers. Traditionally, researchers utilized statistical analysis and modeling frameworks to digest people's mobility patterns and forecast their behavior in the future mostly based on travel surveys and questionnaires from a limited sample size but with a good depth of information (3, 4). Obtaining such datasets are costly and the cumbersome procedure of collecting such datasets makes the traditional source of transportation data not reflect the real-world observations in a timely manner (5). In addition to these shortcomings, the low sample penetration rate, limited period of data collecting, and underreported trips are among the other issues of the traditional data sources which deterred the progress of mobility behavior analysis to some extent.

With the emergence of new technologies, novel sources of data including mobile device location data (MDLD) become accessible to transportation researchers to supplement travel surveys or substitute them in the past two decades. The mobile device location data includes cell phone network location data, GPS devices, and smart mobile phones. Figure 1 shows different types of technologies used as MDLD.

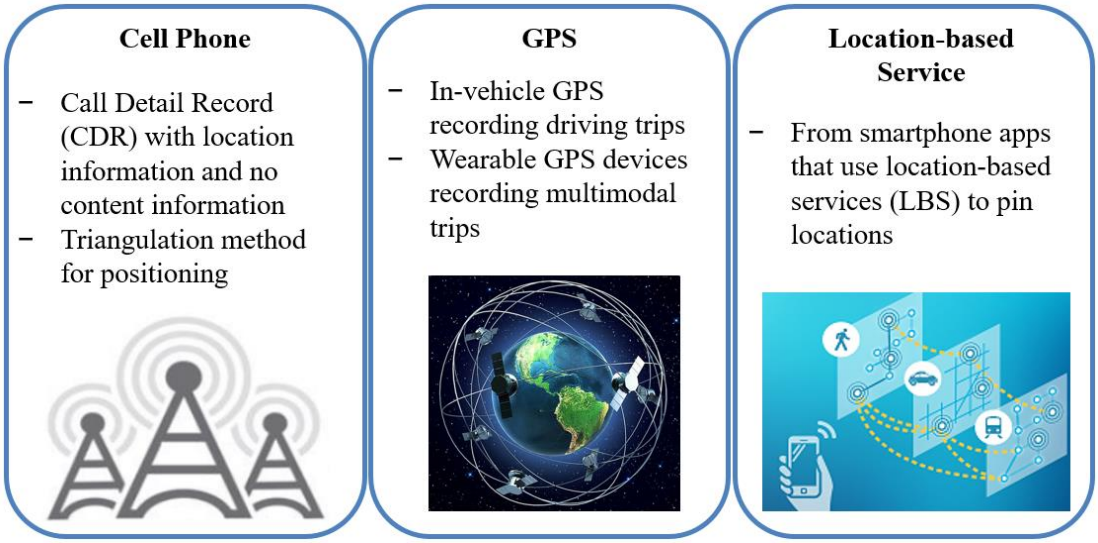


Figure 1. Sources of mobile device location data

As shown in Figure 1, devices as ordinary as cell phones without location services technology as well as dedicated GPS devices in vehicles can contribute to the MDLD data collection. A typical MDLD consists of several core elements including anonymized device ID (either temporary or persistent), the location coordinates of the device, and the timestamp of the event. In addition to these core attributes, location accuracy measurements, speed, and regional time zone offset are among other common features of the data.

Several key advantages of the MDLD over the traditional data sources have attracted researchers' attention in recent years. The continuous recording of device movements passively and objectively, the unprecedented population coverage, and their lower costs compared to traditional surveys are among the most notable features of the MDLD. These advantages encouraged researchers to employ MDLD as a

complementary or even stand-alone data source for the human mobility analysis, including estimation of aggregate-level traffic and travel patterns such as OD tables (6, 7), spatio-temporal human activity pattern analysis (8), Monitoring, modeling, and management of human mobility behavior during extreme conditions such as natural disaster(9), and modeling the human interactions in several contexts including the spread of the disease (10).

Despite all the merits that MDLD possesses, there are several limitations associated with the passively collected data that need to be carefully handled in order to fully exploit this technology. First, MDLD provides no information related to the device owner to preserve the privacy of the data subjects. In addition to no socio-demographic and behavior semantics of the devices, analyzing travel behavior from the MDLD also requires certain data processing and methodological frameworks to extract mobility behavior from the rich spatio-temporal trajectory information of each device.

The increasing interest in leveraging MDLD for different applications and the abovementioned limitations of this data, generate the motivations for this research.

1.2.Objectives

This study has three key objectives. The first objective is to develop a comprehensive framework to analyze and derive the human mobility pattern from MDLD. To achieve this goal, a set of advanced computational algorithms and machine learning methods have been developed to add context to the MDLD and provide necessary

information for further investigating human mobility behaviors through these datasets.

Throughout the Coronavirus disease 2019 (COVID-19) pandemic, the importance of informed decision-making becomes more and more apparent. Therefore, as a part of this study, a framework has been developed to provide near real-time information on the mobility pattern of the communities. Based on the mobility behavior information, a new metric has been introduced to measure the social distancing practices in the communities and to provide more insights on how closely mobility behavior and the outbreak of COVID-19 are linked together.

Lastly, this study seeks to develop a novel framework to investigate the evacuation behavior of individuals during a natural disaster. With the ubiquitous coverage of the MDLD, this study tries to develop a framework to construct the evacuation evolution patterns. In addition to constructing the evacuation decisions, historical mobility behaviors of the device owners were analyzed to further investigate the determinants of the evacuation decision-making procedure based on the revealed mobility characteristics of people.

1.3.Contributions

The first contribution of this study is to introduce a set of new computational algorithms and machine learning methods to extract mobility behavior from MDLD. For this purpose, this study first introduces a set of data preprocessing methods to clean the raw MDLD data. Next, to infer the significant activity location of devices, a

framework has been developed to detect the home and work locations. For the devices with the identified home location, a novel tour-based algorithm is introduced to identify both tours and trips from MDLD. To further investigate the characteristics of the trips, a machine learning method is introduced to detect the mode of travel.

The second contribution of this study is to employ the developed algorithms on a large-scale mobile device location dataset to provide near real-time insights during the COVID-19 pandemic. Non-pharmaceutical interventions (NPI) are considered one of the most effective strategies during the COVID-19 pandemic, especially before the emergence of the vaccines. Mobility restrictions play an important role in containing the virus spread and therefore, understanding how people react to control measures become increasingly important for the decision-makers. To assess the mobility behavior of the communities, this study developed a framework to measure the mobility of the people and constructs a new metric, the social distancing index (SDI), to formulate different aspects of mobility into a single metric that captures the essence of the mobility behaviors related to social distancing practices in different communities.

The third contribution of this dissertation is to build upon the developed algorithms to analyze the individuals' evacuation behavior during the course of a natural disaster such as a hurricane. Complexities of human decision-making procedures and lack of timely data in these situations make the management and planning of the evacuation operations more challenging. To address this need, this study introduces a novel framework to construct the evacuation decisions of people including whether they

evacuate, the departure time and reentry time of evacuations, and their destination choices. Furthermore, this study investigates the determinants of evacuation decisions using statistical models.

1.4.Organizations

The second chapter is dedicated to a comprehensive literature review covering the evolution of state-of-art computational algorithms and imputation methods based on MDLD. Chapter 3 describes the features of the MDLD data along with preprocessing and data cleaning steps needed for the MDLD. In chapter 4, the developed computational algorithms and machine learning methods are introduced. First, the significant activity location identification is described followed by the description of tour based trip identification algorithm. At the end of this chapter, the details of the proposed travel mode imputation algorithm are presented. Chapter 5 shows the MDLD data and discussed algorithms in chapter 4 in action for assessing human mobility during the COVID-19 pandemic. In this chapter, the development of the social distancing index is described. People's mobility behaviors are evaluated based on SDI and the relationships between government orders and the severity of the virus outbreak with the people's behavior are assessed.

In Chapter 6, a novel framework for constructing evacuation evolution patterns is introduced. The results from implementing the framework on the MDLD are summarized in this chapter. Further investigations on how the historical mobility behavior of individuals would impact their decisions toward evacuation are also conducted in the latter part of the chapter by developing two binomial logit choice

models. Finally, the summary of conclusions and remarks for future works are presented in Chapter 7.

Chapter 2: Literature Review

In this chapter, a comprehensive literature review and practice scan have been conducted to cover various topics that are discussed in this dissertation. I grouped the previous research efforts into four subsections. First I summarized the evolution of mobile device location data. In the second part, I reviewed the efforts conducted to extract trip- and device-level information from different types of MDLD. Then, I presented the studies that investigated the importance of mobility behavior in the outbreak of disease. Lastly, this chapter ended with reviewing the studies utilizing MDLD data for evacuation behavior analysis.

2.1. Evolution of Mobile Device Location Data

The earliest attempts to utilize MDLD in the transportation domain started at the end of the last century. In the beginning, the Global Position System (GPS) data loggers were used to collect the longitudinal location data from the survey respondents in order to enhance the quality of the travel diaries (11). The early generation of the GPS data logger required a steady electricity flow and was designed to be implemented in vehicles only using vehicle batteries. The Lexington Area Travel Data was the first survey that utilized in-vehicle GPS technology and proved that the collected GPS data could successfully supplement the traditional approach of collecting manual input from the survey subjects. The in-vehicle GPS data collection has shown to significantly improve the spatiotemporal accuracy of travel records in the survey by capturing the origin and destination of the trips as well as the start time and end time of the trips by collecting the vehicle location second by second while the vehicle is on

(12-17). It has also been shown that in-vehicle GPS data could help to mitigate the issue of underreporting trips and misreporting the trip mileage and travel time estimates (18). The limitation of early GPS technology was that it could only capture the movement of vehicles. As GPS technology improved over the year, wearable and handheld GPS devices helped to record trips made by other modes of transportation by allowing the survey respondents to carry the device. The wearable GPS technology was widely used in travel surveys throughout the past decade (19-21). As travel surveys used both GPS technologies more commonly, several shortcomings remain unresolved including the possibility that users may forget to carry the wearable GPS devices or may consider carrying the device burden and the fact that for some devices, the trip information verification interface is not provided.

As data collection through dedicated GPS devices gets more attention in transportation domains, several research studies investigated the means to extract travel information from the GPS data systematically. Shen and Stopher (2014) revisited methods used for GPS data processing through a review paper (22). They summarized the methodological efforts on GPS data processing for travel survey use cases into three categories: (1) trip/segment identification, (2) travel mode detection, and (3) trip purpose imputation.

In addition to the studies that focused on complementing or replacing the travel surveys (23-25), the GPS data has been utilized for other transportation applications as well. Schonfelder et al. (2002) investigated the feasibility of leveraging longitudinal GPS data to analyze travel behavior. The study used GPS data obtained

from about 400 private and commercial vehicles over the period of two years (26). Papinski et al. (2009) explored the route choice decision-making process by comparing the planned route choice of 31 individuals in Ontario, Canada with their taken route choice observed by a person-based GPS device (27). As the in-vehicle GPS devices become more popular in everyday cars, some private-sector data vendors aggregate such data to provide travel statistics such as travel time, travel speed, link volume estimates, and origin-destination patterns (28). Several scientific reports have assessed the validity of the travel metrics estimated based on these datasets (29, 30).

Since a new generation of mobile devices including mobile phones, smartphones, and tablets, have gained popularity in the past two decades, a new opportunity arises to investigate the human mobility pattern in a more practical approach. The first generation of mobile phone location data was generated using the communications between cellphones and cellular towers (31) based on two different approaches: (1) Call detail record (CDR) data also called event-driven mobile phone data provides details of phone calls and messages including the user id of both sender and receiver, the type of the telecommunication transaction, duration of the transaction, timestamp, and the cell tower ID(s); (2) Network driven mobile phone data that is mainly used by the network carriers to monitor the loads on cell towers or a group of towers named Location Area (LA) to optimize their services (32). In both approaches, the location information is recorded either based on the location of the tower which makes the location accuracy dependent on the density of cellular towers, or in a more precise approach using triangulation algorithms which provide the accuracy of 200 to 300

meters on average (7). Both types of network-based datasets have been used widely to study human mobility patterns in the past two decades. Gonzalez et al. (2008) employed two sets of CDR data to understand human mobility patterns at the individual level (1). In their study, they used CDR data composed of six months of records from 100,000 anonymous individuals selected randomly from a dataset of more than 6 million mobile phone users along with a second dataset that records the location of 206 mobile phone users for every two hours in an entire week. Further studies have been conducted to continue the exploration of human mobility behaviors using a similar dataset (33-38). The CDR datasets are also applied to other research domains such as social network analysis, residential location and population estimation, and predicting socioeconomic levels (39-41). Despite a high penetration rate in the CDR, the data has limitations on both spatial and temporal regards. The spatial accuracy is either confined by the cell towers' density in the network or the accuracy of the triangulation methods. The temporal frequency of observations are also limited by the frequency of the communication transactions such as call and messages.

Location-based service (LBS) data is another source of MDLD which collects spatial and temporal information when a mobile device application updates the device's location by using the most accurate sensor among the existing sensors such as embedded GPS sensor, Bluetooth, Wi-Fi, or cell tower (42, 43). Compared to the CDR, the LBS data possess a higher location accuracy and therefore provide invaluable location information to analyze the individual-level mobility pattern (44). The technology has been used in various transportation-related applications recently.

Resource System Group (RSG) has conducted a smartphone-enhanced travel survey using a mobile application developed by their team, rMove (45). AirSage developed a traffic platform based on LBS data which estimates traffic characteristics of the vehicle movements such as traffic flow, speed, and congestion along with the road user sociodemographic information (46).

In brief, the MDLD sources used in the transportation field are different in several aspects including spatiotemporal coverage of population and their mobility, data quality, e.g. spatial accuracy and location recording interval (LRI), and ease of access to the data (47, 48). The GPS data has the highest horizontal location accuracy (e.g, 10 meters) and the lowest LRI (usually 1 second) while its population coverage is usually very limited and thus cannot represent the mobility behaviors of the entire population. The cellular and LBS data have significantly higher spatio-temporal coverage compared to the GPS data due to the large penetration rate of cellphone and smart mobile devices. However, the data is limited to the spatio-temporal attributes and the LRI for both datasets is high and biased toward users that have more interactions with their devices.

2.2. Extracting Device- and Trip-level Information from MDLD

As the MDLD becomes more accessible to researchers and along with the new developments in the technology, many studies have investigated the extraction of trip information from the raw MDLD. Gong et al. (2014) summarized the methodological attempts conducted to derive personal trip information from GPS data (49). Their reviews included four aspects of the data processing to extract reliable trip

information including trip identification, trip mode imputation, trip purpose detection, and data error recognition that may influence the algorithms. To accurately obtain the trip ends, the first set of algorithms developed used the rule-based trip identification methods that mainly relied on designed rules and corresponding parameters based on the domain knowledge. The rules consider the location data either point by point or several consecutive points at the same time to examine the status of the points whether they are dynamic or stationary. The attributes used in the rule-base models are mostly considering the dwell time, speed, and distance (50-60). Recently, supervised learning machine learning methods are also utilized to supplement the rule-based models to classify the sightings as moving or static (61-63). Unsupervised learning methods such as spatiotemporal clustering algorithms have also been employed for trip end detection. Yao et al. utilized a spatiotemporal clustering method with three layers of optimization models to identify trip ends (64).

With the emergence of the LBS data, additional attempts have been made to identify trips. Wang et al. introduced the "Divide, Conquer and Integrate (DCI)" framework to extract trip ends from multi-sourced data to analyze mobility patterns (44). In their proposed framework, they combined a rule-based algorithm with an incremental clustering method to handle the LBS data with bi-modal nature.

After trip identification, imputation of non-observable attributes is important in order to add context to the identified trips. Significant activity locations such as home and work, trip mode, and trip purpose are among the most important missing attributes.

Home and work location identification are developed based on activity location identification methods. In CDR datasets, as the location records mainly correspond to the cell tower, the area covered by observed cell towers with specific conditions are considered as the significant activity locations. However, for the datasets providing location sightings such as LBS data, the latitude and longitude of each sighting are recorded. Therefore, to analyze the significant activity locations, clustering algorithms have been employed to aggregate static sightings and to identify the home and work area. The algorithms developed to identify the significant activity locations can be categorized into seven classes: threshold-based methods, supervised machine learning, distance-based clustering, model-based clustering, incremental along with K-means clustering, density-based clustering, bi-level modeling framework, and agglomerative clustering approach. Wolf et al. developed a spatial and temporal threshold-based method to detect moving and non-moving sightings by checking all pairs of consecutive points using GPS data (15). Yang et al. and Zhou et al. trained supervised machine learning models to detect static and moving sightings by constructing a feature set from their training datasets (63, 65). Ye et al. and Calabrese et al. investigated a distance-based clustering algorithm by detecting significant stops as a group of consecutive location points that the maximum distance between any pair of points is not larger than the distance threshold and the dwell time is not smaller than the temporal threshold (66, 67). Chen et al. explored the model-based clustering approach to detect significant stops using a Gaussian Mixture Model (68). Wong and Chen developed an incremental approach along with the k-mean clustering method to cluster sightings based on the distance threshold. After identifying clusters they used

a duration threshold in a later step to detect activity locations. The two thresholds were found by trial and error in their investigation (43). Unsupervised machine learning algorithms such as density-based clustering methods have also been investigated to identify the activity location. These sets of algorithms require the number of minimum points and spatial distances to form the cluster. These two parameters are usually selected via trial and error or observations from the raw trajectories (69-72). Wang et al. introduced a bi-level modeling approach by dividing the dataset into two subsets based on their quality. They applied the distance-based clustering algorithm to the high-quality subset and employed the incremental clustering approach to the low-quality subset. The two subsets were integrated by the spatiotemporal relationship at the end (44). The agglomerative clustering method has also been used to complement the previous methods. In this approach, the algorithm consolidates activity locations that are spatially close to each other but may be far away in time (73).

Once the significant activity locations are identified, the activity type such as home and work location should be imputed for each place. The behavior-based and context-based methods are among the most used approaches that have been developed for activity type inference (42). The behavior-based approach classifies the home and work location based on the visiting frequency of the place, the dwell time of each activity location, and the time of day pattern observed in each location (7, 74). On the other hand, the context-based approach utilizes features of the location mainly including the land use type and nearby point of interest (POI) to infer the activity types with predefined empirical rules (75-78).

The behavioral approach has been considered the most widely used method to identify daily life centers such as home and work locations. To determine the daily life centers, Flamm and Kaufmann proposed the criteria of individuals spending at least 20 percent of their time based on their investigation on the Moby drive dataset that contains six-week survey period information (79). Calabrese et al. proposed grids of 500*500 meters (1640.42*1640.42 ft) to label the activity location. They considered grids with most night-time observation, the period from 6 p.m. to 8 a.m., as the home location. The work locations were similarly identified as the most frequent observed grid on weekdays between 8 a.m. and 10 a.m. They validated their results against the Census Transportation Planning Products (CTPP) (80). In addition to the data-driven approaches, supervised learning methods have also been considered in identifying the activity location. Isaacman et al. developed a feature set of five observable attributes and derived 15 factors by ranking and calculating the percentage of the observable attributes. A logistic regression model has been trained based on the feature of 15 factors using a labeled dataset collected from 18 volunteers (81).

After identifying trips, the mode of the trip is another important aspect of the mobility behavior that needs to be imputed. Travel mode imputation can be categorized into two approaches mainly: trip-based approach; and segment/point-based approach. The trip-based approach is based on the already identified trips to detect a single travel mode for the entire trip while in the segment/point-based approach, the travel mode for each segment or point is being imputed separately (48). Then the segments/points with the same travel mode are merged to form a trip with a single mode. To

distinguish the mode, both approaches have used similar features. Table 1 summarized the feature sets used in the travel mode imputation previously (82).

Table 1. Literature review on travel mode detection methods

Author	LRI	Model*	Main Features	Modes	Acc.
Gong et al. 2012	/	Rules	Speed, Acceleration, Transit Stations, Transit Network	Drive, Train, Bus, Walk, Bike, Static	82.6%
Stenneth et al. 2011	30 s	RF	Speed, Acceleration, Heading change, Bus location, Transit Network	Drive, Bus, Train, Walk, Bike, Static	93.7%
Bruunauer et al. 2013	1-10 s	MLP	Speed, Acceleration, Bendiness	Drive, Bus, Train, Walk, Bike	92.0%
Xiao et al. 2015	1 s	BN	Speed, Acceleration, Trip Distance	Drive Bus, Walk, Bike, E-Bike	92.0%
Nitsche et al. 2014	1 s	DHMM	Speed, Acceleration, Direction	Drive, Bus, Motorcycle, Train, Tram, Subway, Walk, Bike	65% - 95%
Dabiri and Heaslip. 2018.	1-5 s	CNN	Speed, Acceleration, Jerk, Bearing Rate	Drive, Bus, Train, Walk, Bike	84.8%
Bachir et al. 2019	/	BI	Road and Rail Trip Counts	Road, Rail	-
Vaughan et al. 2020	/	DNN	Speed, Trip Distance, Land Use, Time of Day	Drive, Bus, Active (Walk, Bike)	87%
Burkhard et al. 2020	1 s subsampled to 5 min	KNN, RF etc.	Speed, Public Transport Stops and Lines	Drive, Train, Tram, Bus, Walk, Bike	-
Breyer et al. 2021	/	KNN etc.	Road and Train Route Geometry	Road, Train	95.5%

** RF: Random Forest; MLP: Multi-Layer Perceptron; BN: Bayesian Network; DHMM: Discrete Hidden Markov Model; CNN: Convolutional neural Network; BI: Bayesian Inference; DNN: Deep Neural Network.*

Based on the literature review conducted by Huang et al. and Burkhard et al (47, 48), speed and acceleration are among the typical features of mode imputation studies (48, 58, 83-91). Especially, when the location recording interval (LRI) is less than 10s, the speed variation, and acceleration features are more important to differentiate between various travel modes. On the other hand, when the LRI becomes relatively higher (e.g. more than 30 seconds), the importance of additional features is becoming higher to maintain the same level of accuracy. Real-time transit information (83), multimodal transportation network (48, 58, 83, 92), and socio-demographic information (88, 91) are among additional features that have been investigated in past studies.

2.3. Impact of Mobility Behavior during Pandemic

As MDLD gain popularity in studying human mobility behavior in recent years, the application of this data source has been proven to be a great asset for decision-makers amid the current COVID-19 pandemic.

The effect of mobility patterns and non-pharmaceutical interventions such as social distancing has been well-studied for preventing virus spread (93-95). Empirical analysis utilizing airline travel revealed the significant influence of international air travel on the progress of influenza outbreaks, as well as the impacts of domestic air travel on the evolution of disease spread across the United States (94). Later on, studies utilized more comprehensive mobility data to investigate the influence of mobility patterns and travel restrictions on containing the epidemic spread (10, 95). As one of the major non-pharmaceutical interventions, social distancing is considered

an effective way to reduce COVID-19 infections, especially in the pre-vaccine period. Researchers have highlighted the important role of social distancing in disease prevention through modeling and simulation (96-99). The simulation models assume a level of compliance based on the generated synthetic population (100), estimated contact patterns using survey data (101, 102), or collect people's behavior reactions through dedicated surveys (103). Furthermore, artificial intelligence (AI) techniques, along with big data, have also been largely applied in several different aspects of managing the COVID-19 pandemic, such as early detection and diagnosis, monitoring the treatment, contact tracing of individuals, and projection of case and mortality (104, 105). The lack of timely contributions from real-world observations became apparent at the beginning of the pandemic as the studies tried to model the evolution of the outbreak. Many companies such as Google, Apple, and Cuebiq started to produce valuable information about mobility and economic trends (106-108). These analyses mainly focus on a single indicator of the mobility aspect such as distance traveled or visitations to various business sectors.

2.4. Disaster Evacuation Behavior Analysis

There is a wide range of research studies focused on various types of disasters. In this section, I mainly focus on evacuation behavior studies. Several studies reviewed the literature on evacuation behavior (109), evacuation modeling (110), and common transportation practices during evacuation (111).

Many studies focused on a specific disaster or set of disasters to analyze the important factors in evacuation behavior, evaluate the disaster planning and preparation, or

assess disaster management and logistics. Collier et al. (2019) studied major transportation and logistics issues and summarized lessons learned from the two major hurricanes in the U.S., Hurricane Katrina, and Hurricane Harvey. In their study, they provided recommendations for future hurricanes considering the evacuation planning, information provision, infrastructure management, and disaster preparation aspects (112). Simulation models are widely used in disaster planning and management studies (113-119). Feng and Lin (2019) used a hurricane-prediction demand generation model in a fast agent-based modeling framework calibrated with traffic observations to study evacuation during Hurricane Irma (116).

The evacuation behavior studies traditionally relied on surveys (120-124). These post-hurricane surveys are traditionally used to collect information regarding various evacuation decisions i.e., evacuating or not, departure time of the evacuation, destination choice, primary travel mode used for the evacuation, route choice, and reentry time decisions (125, 126). For instance, Kontou et al. collected telephone survey data from commuters affected by Hurricane Sandy and estimated a hazard-based model to identify the parameters that affect the duration of commute behavior changes (124). Wong et al. collected an online survey from individuals impacted by Hurricane Irma and studied their evacuation behavior. In their study, their summarized descriptive statistics and developed statistical models for various decisions made during Hurricane Irma (120). Although these surveys are usually rich in terms of recording evacuee's decisions and revealing their preferences during the disaster, such surveys are costly, implemented for a small number of respondents, time-consuming, and not capable of providing real-time information.

With the increasing availability and popularity of big data, new approaches are now available for studying long-lasting questions. Robinson et al. identified two main challenges in studying disaster evacuation; the first was the complexity of human behavior and the second was data deficiency for traffic information and household decisions (127). Both issues can be resolved to some extent by utilizing MDLD. MDLD does not provide detailed individual-level information, but with its significant sample size, and proper data processing, it can reveal valuable information for many critical evacuation-related behaviors. Besides the larger sample size, MDLD has other advantages over traditional surveys. First, the phenomenon known as the *observer effect* (128), which suggests that individuals may modify their behavior when being observed or studied can be addressed by MDLD due to its passively collected nature. Passively collected data capture the normal behavior of subjects, free of any study-related observer error. The second is related to known survey design errors such as sampling error, measurement error, response error (129), and survey response biases (130). Even though MDLD may have its own biases (such as bias toward higher-income populations) and errors (such as inaccurate sightings), it records the actual behavior of subjects, not recalled or stated behavior. The third aspect is specific to disaster-related surveys. Surveying individuals about traumatic events may sometimes be undesired for the respondents. Passive data collection does not put any emotional burden on the respondents.

Considering all the advantages of the MDLD, more recent studies are taking advantage of big data for evacuation behavior studies. Social media data was among the first MDLDs that has been utilized in the evacuation analysis. Kumar and

Ukkusuri (2018) utilized geo-tagged tweets from New York City at the time of Hurricane Sandy to study the evacuation behavior of affected residents (131). Their study showed a strong relationship between social connectivity and the decision to evacuate. Roy and Hasan (2021) collected Twitter data related to Hurricane Irma and developed a Hidden Markov framework to model the dynamics of hurricane evacuation and infer evacuation decisions (132). Wang and Taylor (2014) also used Twitter data to study the correlation between movement patterns under steady-state and perturbed state during Hurricane Sandy (133).

Compared to the social media generated data, LBS data has a higher penetration rate and smaller demographic biases. However, the application of LBS data in evacuation studies remains very limited. Yabe et al. (2020) collected LBS data for five disastrous events (1.9 million devices in total) to study recovery patterns at the macroscopic population level and showed similarity in recovery patterns of these events despite differences in population characteristics (134).

Chapter 3: Data

The emergence of mobile device location technologies such as cellphone, GPS, and LBS made MDLD a prominent asset in various application areas including human mobility behavior analysis. This section describes the methodology for assessing raw location data quality. A typical MDLD record from LBS technology contains information about timestamp, anonymized device ID, location of the device (latitude and longitude coordinates), a measure of spatial accuracy. In some cases, additional information such as the device operating system (OS) and time zone offset of the position of the device are also provided. A synthetic sample of data is provided in Table 2 to demonstrate the raw data. Entries presented in Table 2 are modified to preserve privacy.

Table 2. A synthetic sample of LBS data

Timestamp	Device ID	Device Type	Latitude	Longitude	Location Accuracy (m)	Time Zone Offset
1504068337	e07941996a2ffd303021914	1	28.4302	-81.6065	5	-14400
1504068342	e07941996a2ffd303021914	1	28.4303	-81.6053	25	-14400
1504068351	e07941996a2ffd303021914	1	28.4302	-81.6042	5	-14400
1504068360	e07941996a2ffd303021914	1	28.4305	-81.6046	100	-14400
1505096982	F258069021658ssd132548e	0	28.4313	-81.6037	5	-14400

In some cases in the raw data, because of privacy protection, the location information may be reported in an aggregated or transformed form.

3.1. Data Cleaning and Preprocessing

Although mobile device location datasets are rich in terms of spatio-temporal characteristics, certain treatments and data cleaning steps are needed before extracting any information from the data. Removing outliers, checking for potential consistency issues in the data (e.g. unreasonable high-speed records), identifying duplicate observations for the same device, and merging them are among the state-of-practice methods for cleaning raw data and controlling its quality. The data cleaning approach proposed in this study first investigates the four well-known aspects of the data quality assessment framework: consistency, accuracy, completeness, and timeliness (135). To ensure the consistency of the data, certain semantic rules have been defined such as integrity constraints, to be checked through the entire raw data. At this step, all data entries are evaluated to identify observations with invalid values. For example, the latitude and longitude information of a location should follow a reasonable range, so integrity constraint removes all records with invalid entries. The other check is to identify duplicate records to reduce data redundancy and size to facilitate the computational process. Since one device should only be present in no more than one location at the same time, this procedure keeps only one data entry with the highest spatial accuracy at a certain time for one device.

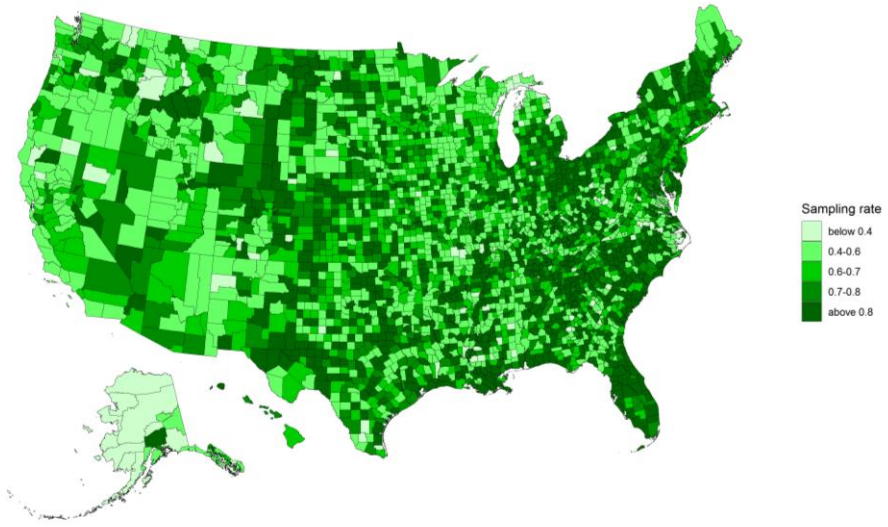
Accuracy is another important dimension of data quality assessment, covering both syntactic and semantic accuracies. The semantic accuracy evaluates the closeness of a value to its real-world observation while syntactic accuracy ensures the closeness of a value to the elements of its corresponding definition domain. In this application, a

spatial accuracy of 50 meters indicates that the device should be within 50 meters of the reported location with a certain confidence interval, for example, 95%. Thus entries with extremely poor spatial accuracy (i.e. location accuracy attribute of higher than 2 miles) are removed from the dataset based on the semantic accuracy rule.

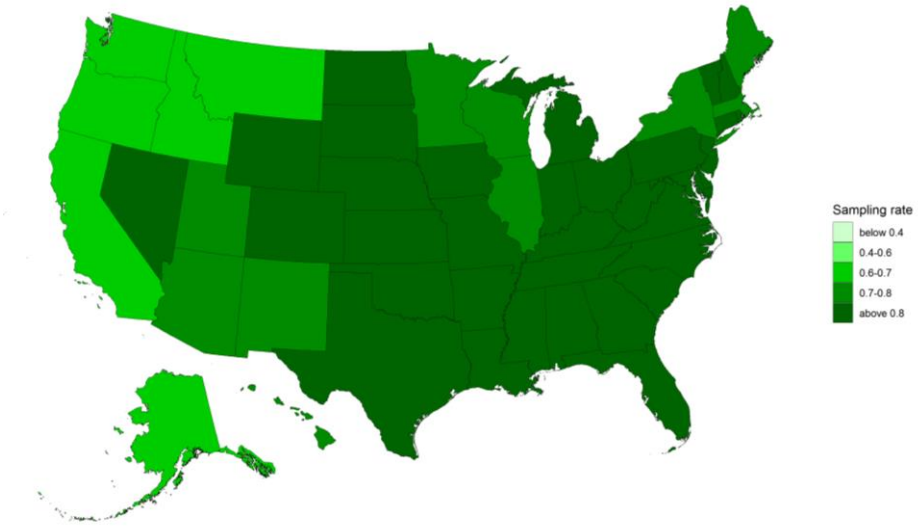
The completeness aspect requires prior knowledge of the actual movement patterns and mobile device usage, which is not available in this application. Therefore, this dimension has not been incorporated into the data cleaning procedure. For the timeliness dimension due to the timely nature of the applications introduced in this dissertation, an attempt is made to consider it by incorporating daily feeds of location in the data pool.

3.2. Data Summary

After conducting data cleaning and preprocessing checks on the raw data, the cleaned data covers more than 270,000,000 Monthly Active Users (MAU) for February 2020 representing movement information across the nation. Figure 2 depicts the coverage of the raw sighting data at different geographical levels.



(a) Device sampling rate at the county level



(b) Device sampling rate at the state level

Figure 2. Device sampling rate for the month of February 2020 (a) at the county level, (b) at the state level

Figure 3 demonstrates the heatmap of sighting density for the continental U.S (136).

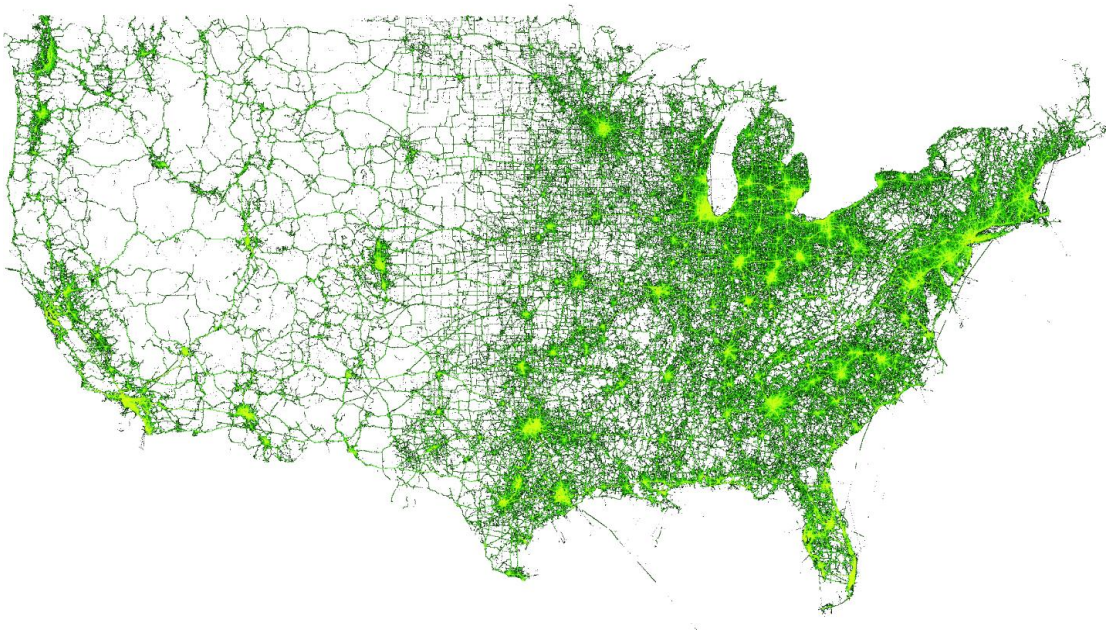


Figure 3. The density map of anonymized location data across the nation (brighter shades represents a higher density of sightings)

Chapter 4: Deducing Device- and Trip-Level Information

This section describes the methodological advances this dissertation proposes to enhance extracting device- and trip-level information from large-scale mobile device location data sources.

4.1. Home and Work Location Identification

Due to privacy protection, the mobile device location datasets are generally anonymized and do not contain any personally identifiable information (PII). Therefore, researchers should develop home and work location identification algorithms to add context to the extracted information from the MDLD. In this dissertation, a behavior-based method has been proposed that evaluates the temporal patterns of places observed for every device and ranks the frequently visited location to identify the home and work location at a monthly cadence.

To efficiently process the tremendous amount of MDLD, the algorithm utilizes the geohash notion, a public domain geocode system that encodes a geographic location into a short string of letters and digits, to aggregate the latitude and longitudes into candidate clusters for significant activity location. Geohash cell dimensions vary with the latitude of the location. Table 3 summarizes geohash sizes at the equator.

Table 3. Geohash cell dimensions at the equator

Geohash string length	Width	Height	Geohash string length	Width	Height
1	5,009.4 km	4,992.6 km	**7**	152.9 m	152.4 m
2	1,252.3 km	624.1 km	**8**	38.2 m	19 m
3	156.5 km	156 km	**9**	4.8 m	4.8 m
4	39.1 km	19.5 km	**10**	1.2 m	59.5 cm
5	4.9 km	4.9 km	**11**	14.9 cm	14.9 cm
6	1.2 km	609.4 m	**12**	3.7 cm	1.9 cm

Considering the location uncertainty of sightings and activities conducted near the home location, the algorithm identifies the significant activity location in a bi-level approach. First, home and work locations are identified at the level-6 geohash to minimize the effect of the noises, and then to derive a more precise representation of the home and work locations, the algorithm searches for the best candidate at level-7 geohash cells within the identified level-6 geohash.

As suggested in the literature, people spend most of their time, especially nighttime, at home and some fixed and regular hours during daytime at the workplace. To determine the nighttime, the time activity pattern from American Time Use Survey (ATUS) has been reviewed. According to 2017, 2018, and 2019 ATUS, more than 80% of full-time and part-time workers, who are observed to visit home at least once during the survey day, stay at home during the 21:00-5:59 period. Therefore, the nighttime window is defined as 21:00-5:59.

Identifying home location at geohash level-6 follows the following steps:

1) Observed on at least $\max\left\{3, integer\left(\frac{Number\ of\ observed\ days}{2}\right)+1\right\}$ days;

2) Observed on average more than h (≥ 2) hours daily;
3) Sort the home candidates by observed number of days, average daily number of observed hours, and average number of hourly sightings;
4) Keep 3 top-ranked home candidates and sort them by observed number of nights, average daily number of observed nighttime hours, and average number of hourly sightings during nighttime;
5) Select the top-ranked level-6 geohash as the home location; in case of need for a tie-breaker, select based on step 3.

The first 2 rules were implemented to ensure the minimum quality needed for keeping a device in our data pool.

Once the home location has been identified at geohash level 6, the best level-7 geohash candidate selects based on the following rules:

1) Filter observations for all corresponding level-7 geohashesh within the identified level-6 home geohash;
2) Sort the level-7 geohash candidates by observed number of days, average daily number of observed hours, and average number of hourly sightings;
3) Keep 3 top-ranked candidates;
4) Sort the home candidates (level-7 geohashes) by observed number of nights, average daily number of observed nighttime hours, and average number of hourly sightings during nighttime;
5) Select the top-ranked level-7 geohash as the home location; in case of need for a tie-breaker, select based on step 2.

The objective of work location identification is to determine an individual's major work location that is not the same as their home location. Therefore, level-6 geohashes that are not one's home geohash have been considered. In addition, the algorithm introduces a temporal similarity ratio on top of the commonly used attributes in behavior-based methods such as the frequency of visits, dwell time, and regularity. The motivation for utilizing the temporal similarity ratio is two-fold. First, since the algorithm is adopting geohash grid-based geocode system instead of a spatial or spatio-temporal cluster of sightings due to computational efficiency, in case a device dwells around the borders of geohash zones, a neighboring geohash zone can record frequent observations. This one or more than one neighboring geohashes – twin zones- could become a competitive candidate for the actual workplace zone in terms of visiting frequency, duration, and regularity. Second, although a minimum commute distance may seem to be an intuitive alternative to address the aforementioned issue, selecting a universal minimum distance may compromise workplaces that are close to one's identified home location. Based on the assumption that one shall commute from home to work and work for consecutive hours before arriving back home, the temporal similarity ratio imposes a condition that home and work location shall not be frequently observed at the same hours.

Hence, the temporal similarity ratio is defined as follows. For all the unique hours when a workplace candidate was observed during the month, i.e. W^i for candidate i, count the number of unique hours overlapping with all the unique hours when the imputed home location was observed H. The ratio between the overlapped

hours and the total number of hours in W^i is then calculated. The ratio, referred to as temporal similarity ratio S, measures the temporal similarity between home and workplace observations. The formula is given as follows.

$$S^i = \frac{|W^i \cap H|}{|W^i|} \quad (1)$$

In an ideal case where the daily location observations are complete for one device with a fixed workplace, the ratio should be $\frac{2}{Number\ of\ daily\ work\ hours}$ considering the departure time of the commute and when the commute time is shorter than one hour, and zero when the commute time is longer than one hour. However, considering that the complete location observation is not available for most of the devices in MDLD throughout the month, imposing a small temporal similarity ratio would lead to exclusion of actual work locations. To address this, the algorithm is designed to favor work candidates with smaller temporal similarity ratios while imposing a maximum temporal similarity ratio threshold to exclude the inefficient large ratios to distinguish between the actual work location and the twin zones of home location.

The algorithm identifies level-6 geohash work location based on the following rules:

1) Observed on at least $max\left\{3, integer\left(\frac{Number\ of\ observed\ workdays}{2}\right) + 1\right\}$ workdays;
2) Observed on average more than W (≥ 2) hours daily;

3) Sort the work candidates by observed number of workdays, average workday number of observed hours, and average workday number of hourly sightings;
4) Keep the three top-ranked candidates
5) Calculate temporal similarity ratio, S, following equation (1);
6) Sort the three work candidates (level-6 geohashes) by similarity ratio in the ascending order;
7) Select the top-ranked level-6 geohash with a similarity ratio smaller than the maximum temporal similarity threshold as the work location.

Once the work location is selected at level-6 geohash, for a more precise representation of work location, the following set of rules are defined to search for the best level-7 geohash candidate among all the level-7 geohashesh within the identified level-6 geohash work location.

1) Start from all the corresponding level-7 geohashes within the level-6 geohash workplace;
2) Sort the level-7 geohash candidates by observed number of workdays, average workday number of observed hours, and average workday number of hourly sightings;
3) Select the top-ranked level-7 geohash as the work location.

There are two major parameters to be calibrated in the introduced algorithm, the minimum observed daily hours for home, $H(\geq 2)$ hours, and workplace, $W(\geq 2)$ hours. To calibrate the H parameter, the Pearson correlation between the county-level number of imputed residents and the population over 16 reported by the

American Community Survey (ACS) (137) is calculated for different values of H (see the dark green line in Figure 4). For workplace calibration, the Pearson correlation between the county-level number of imputed commuters and the number and the number of workers reported by Longitudinal Employer Household Dynamics (LEHD) Origin Destination Employment Statistics (LODES) (138) is calculated for different combinations of H and W (see the black dotted line in Figure 4). Figure 4 implies that increasing the minimum observed hours for home and work leads to a decrease in the Pearson correlation. Therefore, the combination of two for H and two for W is selected to yield the best performance in imputing home and work location identification.

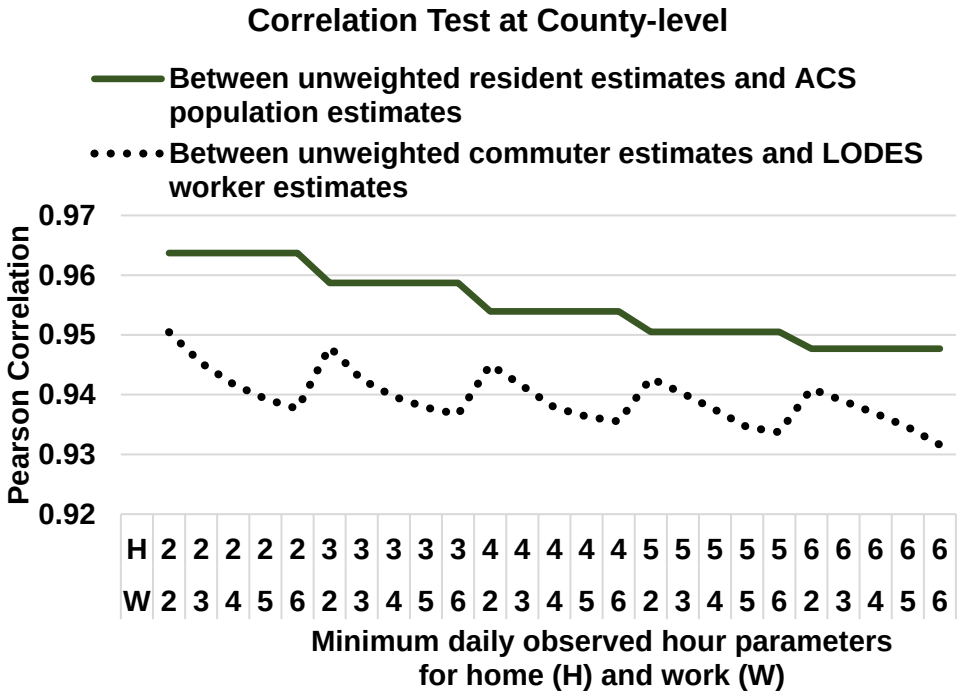


Figure 4. Calibration results for selecting the number of minimum observed hours.

In addition to the minimum observed hours, two reasons lead to selecting the maximum temporal similarity ratio of 0.6. First, the workplace should be observed for at least one specific hour in each visit excluding the home location observations besides the two shared observed hours during the two commute trips (with the consideration of short commute trips and departure time of commutes). Second, a sensitivity analysis regarding the maximum threshold was conducted considering the county-level Pearson correlation between the imputed number of workers and the reported number of workers in LODES (see the dark green line in Figure 5) and the percentage of devices with imputed workplace over devices with identified home (white bars in Figure 5). Figure 5 shows that by increasing the similarity ratio parameter, the Pearson correlation decreases with a platoon between 0.2 and 0.6 while the number of devices with imputed work location increases at a steady pace. Therefore, 0.6 as a similarity ratio balances the tradeoff between Pearson correlation and avoiding failing to identify the work location for many devices.

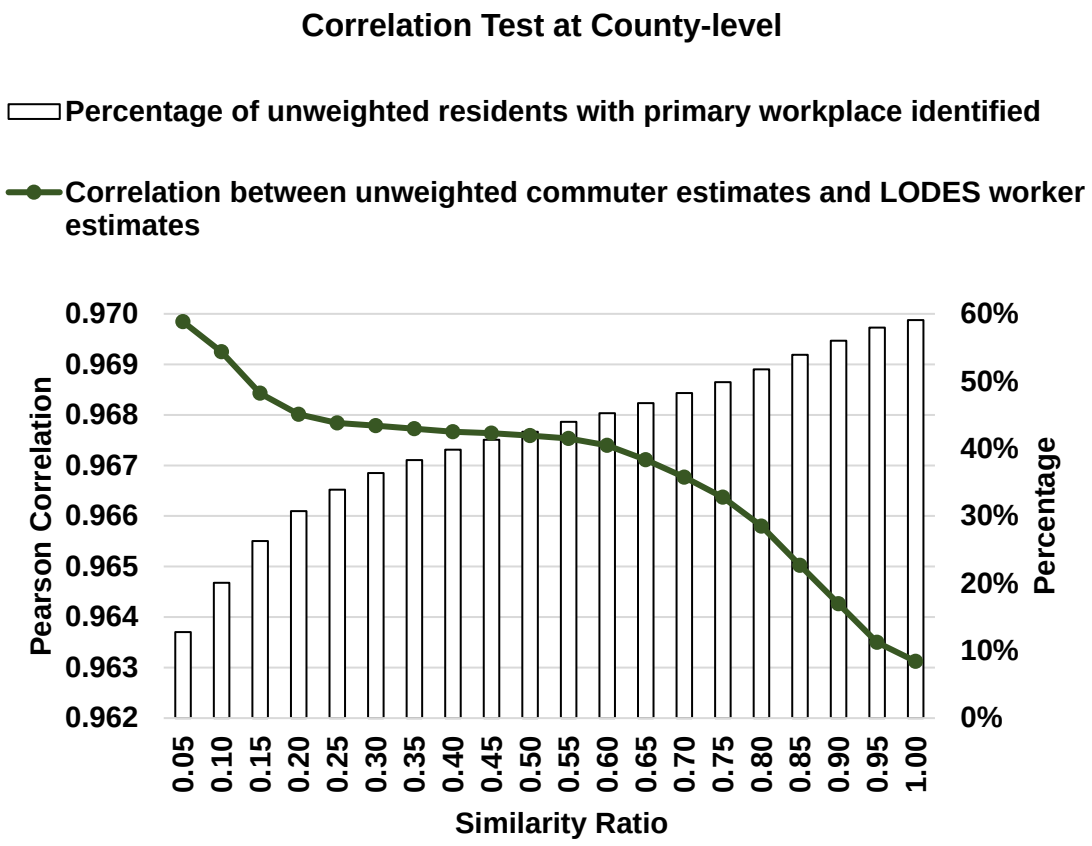


Figure 5. Sensitivity analysis in temporal similarity ratio using MDLD

4.1.1. Comparisons with Alternative Home Identification Algorithms

In addition to the proposed method, this study examines two alternative home identification algorithms and compares their performances using a mobile device location sample dataset. The first algorithm (referred to as the "nighttime method" in the following context) is a widely used state-of-the-practice method, which identifies the home location as the place with the highest observed hours from 6 p.m. to 7 a.m (139). The second alternative is a conservative method (referred to as the "all-day method"). It first applies a strict filter to the home candidates. Each level-7 geohash

candidate must be observed for at least 14 days and at least 60 distinct hours within the study month. Then, it identifies the home location as the level-7 geohash with the highest observed hours. When a tie exists, the level-7 geohash with the most sightings is selected.

The results show that the nighttime method yields the most imputed residents, i.e., 74% of all the devices in the raw data, followed by the proposed method (12%) and the all-day method (8%). Next, the county-level Pearson correlations between the imputed residents and the ACS population are 0.966 for the nighttime method, 0.969 for the proposed method, and 0.962 for the all-day method, where the proposed method slightly outperforms the other two approaches. Moreover, the distances between the home locations imputed from the three methods are calculated and summarized in Table 4. Each column is based on the imputed home locations of the same imputed residents shared by two methods. It can be observed that the discrepancy between the home location starts at 90^{th}-percentile for the nighttime to proposed comparison, and 95^{th}-percentile distances for all three cases are smaller than 1 mile. Although the nighttime method yields similar home locations to the all-day method for their shared imputed residents, the distances between its imputed home locations and the proposed method's home locations are the largest. By jointly considering the sample size reduction, the Pearson correlation to the ground truth population, and the differences in the imputed home locations, the proposed method yields the overall best results.

Table 4. Descriptive statistics on the distances between the imputed home locations

Measure (Miles)	Nighttime to Proposed	Nighttime to All-Day	Proposed to All-Day
Mean	4.46	1.90	1.40
75%	0	0	0
90%	0.07	0.00	0.00
95%	0.85	0.09	0.29
99%	30.09	17.66	17.26
Max	5892.71	5098.57	4972.37

To further dig into the comparisons, Figure 6 shows the scatter plot of the county-level resident estimates between each of the introduced algorithms and the ACS.

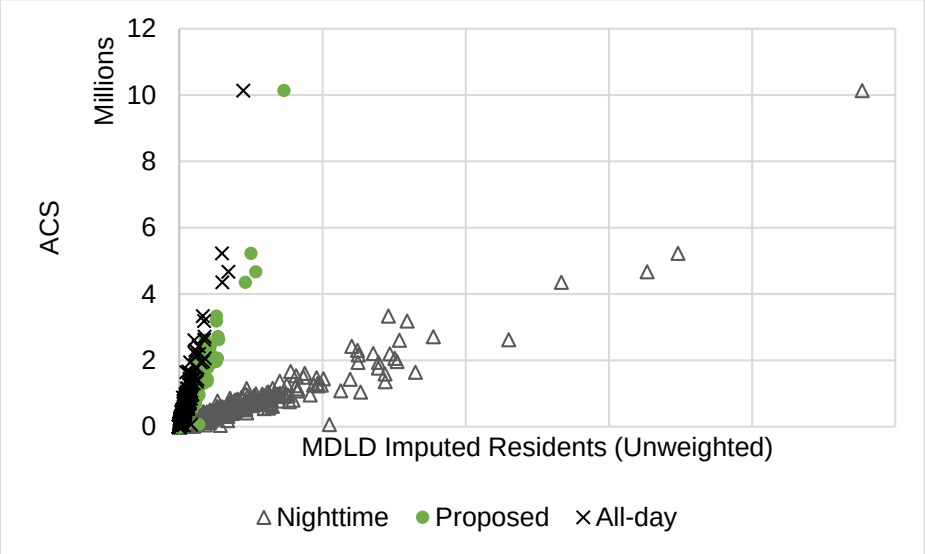
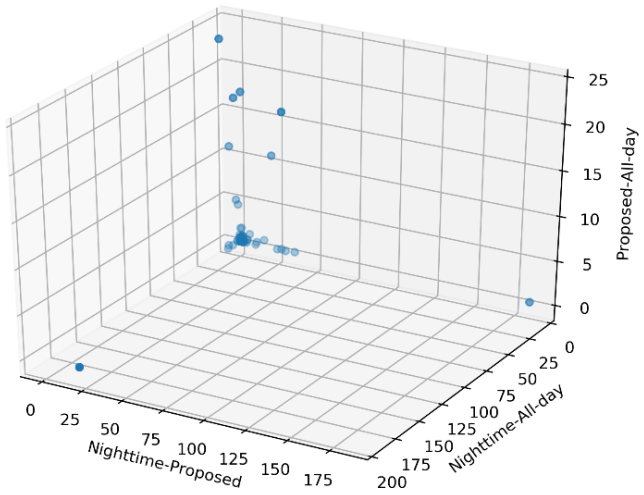


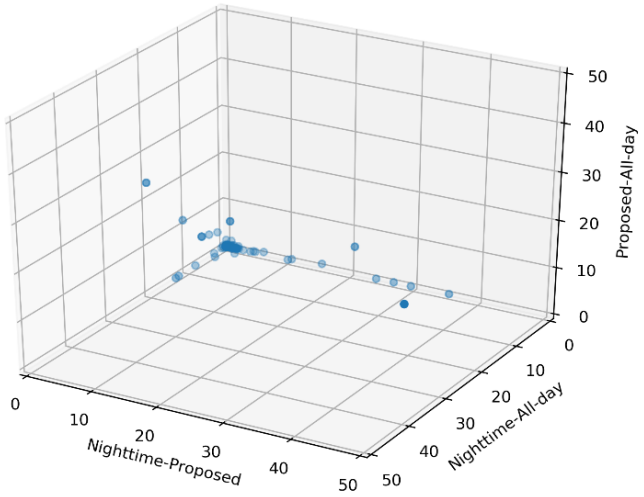
Figure 6. County-level resident estimates from different methods and ACS

Figure 7, Figure 8, and Figure 9 demonstrate the 90th-percentile, 95th-percentile, and 99th-percentile distances between the home locations imputed from each pair of the three introduced algorithms at the county level, respectively. Each figure first

displays the entire three-dimensional scatter plot, followed by a zoom-in plot. All distances are measured in mile.

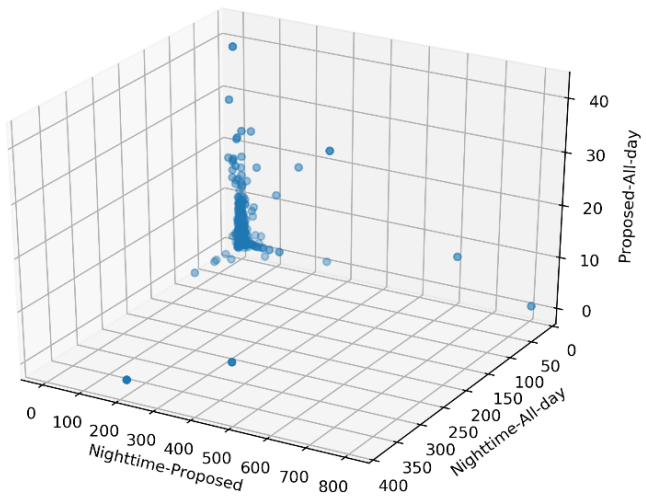


(a) All counties

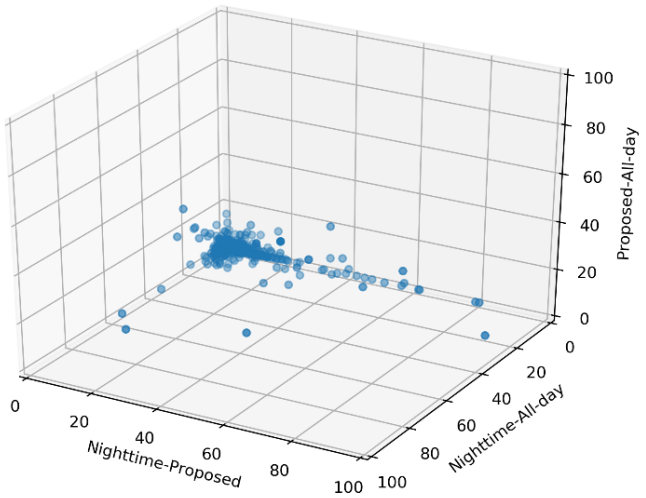


(b) More zoom-in plot

Figure 7. County-level 90th-percentile distances between the imputed home locations

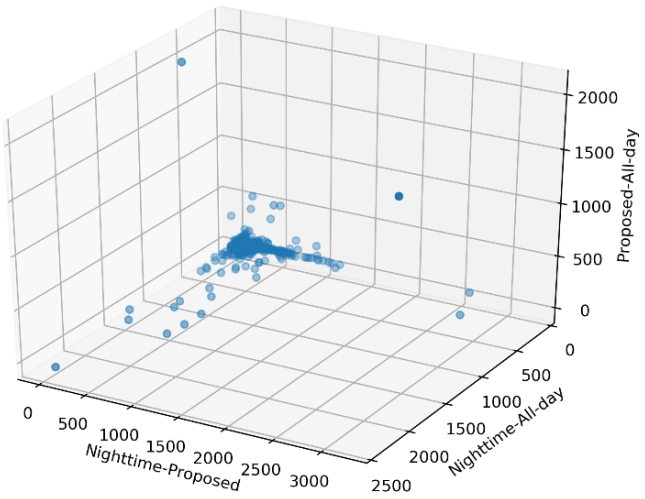


(a) All counties

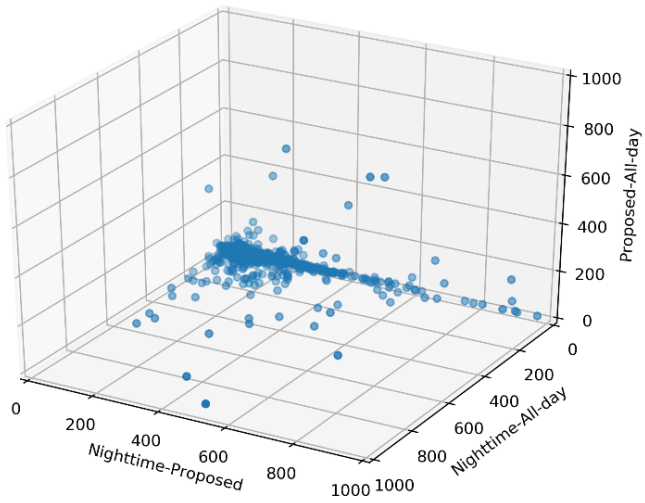


(b) More zoom-in plot

Figure 8. County-level 95th-percentile distances between the imputed home locations



(a) All counties



(b) More zoom-in plot

Figure 9. County-level 99th-percentile distances between the imputed home locations

4.1.2. Home and Work Location Identification Validation

Since the mobile device location dataset used in this research does not contain any ground truth information on the home and work locations, the identified daily life centers are validated against the ground truth population and employment statistics. With the calibrated parameters, the MDLD sample devices are aggregated at the county level based on the imputed home locations for further analysis. The spatial distribution of the unweighted MDLD resident estimates is compared with that of the 2019 ACS 5-year population estimates (137) in Figure 10, which shows similar spatial distributions estimated from MDLD and ACS with a Pearson correlation of 0.970.

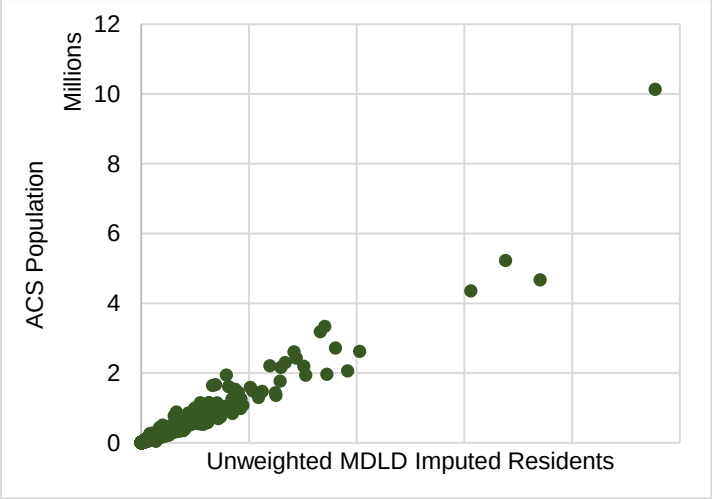


Figure 10. County-level resident estimates comparison between MDLD and ACS

Similarly, the MDLD sample devices with both imputed home and fixed workplaces are considered normal commuters and are aggregated at the county level based on the imputed home locations. The ground truth data from the 2019 LODES estimates

(138) and 2019 ACS 5-year estimates (140) have been adjusted to the 2020 estimates with a national-level population inflation factor of 1.005. The spatial distribution of the unweighted commuter estimate is then compared with the two ground truth datasets in Figure 11. Figure 11 shows similar spatial distributions of unweighted normal commuter estimates from MDLD and ACS with a Pearson correlation of 0.969 and from MDLD and LODES with a Pearson correlation of 0.967.

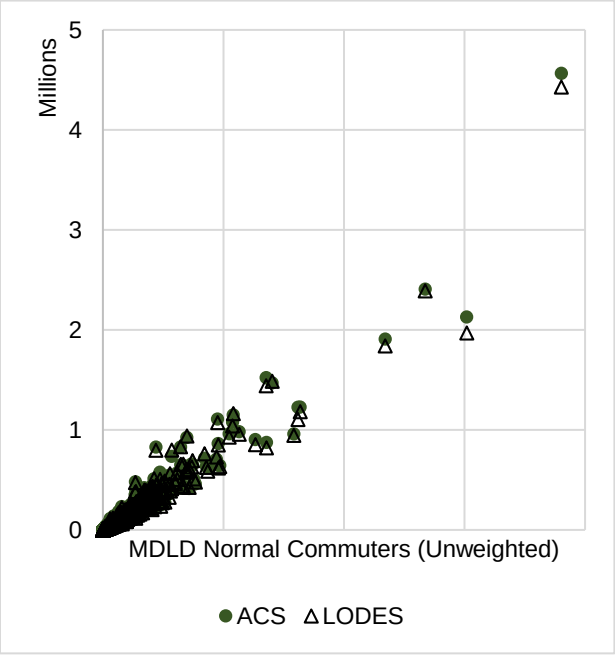


Figure 11. County-level normal commuter estimates from MDLD, ACS, and LODES

Next, the commuting flow estimates from MDLD are validated against 2019 LODES (138) and 2015 ACS 5-year commuting flow estimates (141). Following the spatial resolution of the ACS estimates, the MDLD and LODES estimates are aggregated at the county level from level 7 geohashes and census block groups, respectively. Due to the differences in the data collection and coverage, the two ground truth data sources

can produce distinctive results for some queries (142). Related to the commuting origin and destination (OD) pairs, the two data products have different home and work location definitions. In the ACS data, the work location is provided by the survey respondents as the specific work address during last week. On the other hand, the work location in the LODES data is reported by the employers which can be an administrative address instead of the actual worksite. Meanwhile, the residence location in the LODES data is based on a residence synthesizer and can be outdated if a worker moves during the year. As a result, there are 815,941 unique OD pairs from LODES and only 135,904 unique pairs from ACS. Figure 12 compares the commuting flow estimates for the shared OD pairs between LODES and ACS. It can be observed that the ACS data have higher estimates due to fewer unique pairs.

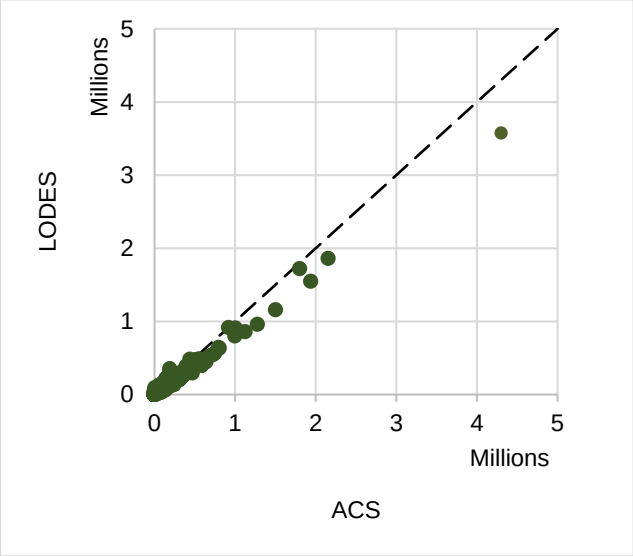


Figure 12. County-level commuting flow estimates from ACS and LODES

In Figure 13, the commuting flow estimates from MDLD are compared to LODES estimates. From MDLD, there are only 120,458 unique OD pairs, which may be due to the fact that the estimates are from a relatively short period of time (January 2020). In addition, the MDLD home and work locations are imputed based on the actual location observations, which are more similar and consistent with the definitions in the ACS data. With all that considerations, a similar trend is observed for the shared OD pairs with a Pearson correlation of 0.951.

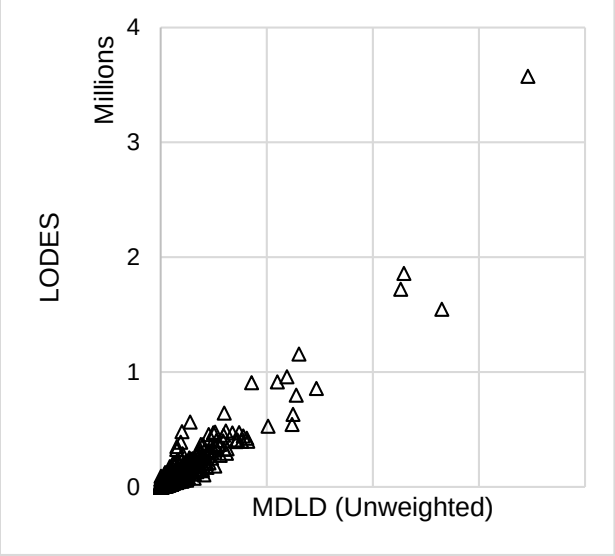


Figure 13. County-level commuting flow estimates from MDLD and LODES

Figure 14 further compares the unweighted commuting flow estimates from MDLD and ACS estimates. It can be observed that the MDLD estimates share a closer pattern with ACS estimates. The Pearson correlation is 0.965 for this comparison.

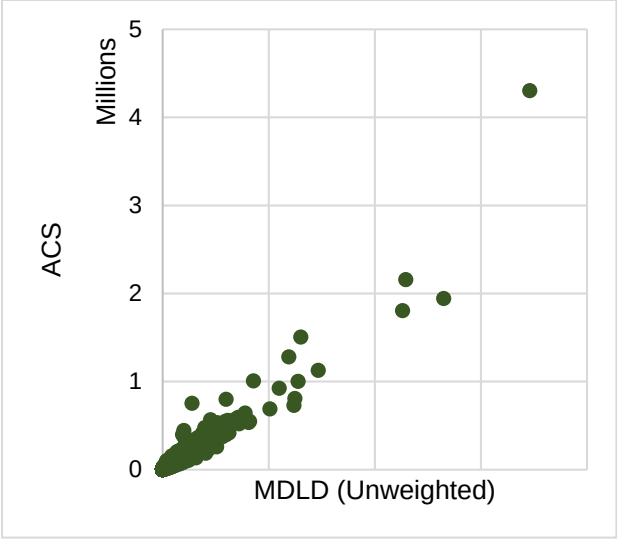


Figure 14. County-level commuting flow estimates from MDLD and ACS

In addition, the commuting distance distributions from the three sources are compared in Figure 15. The commuting distance is calculated as the mileage between the centroids of the home and work counties for consistencies. In general, the MDLD distribution has very similar patterns to the ACS estimates while both of them have higher estimates for shorter distance bands. It suggests that LODES observes more long-distance OD pairs and more long-distance commuters than ACS and MDLD, which can result from the aforementioned definition for home and work locations.

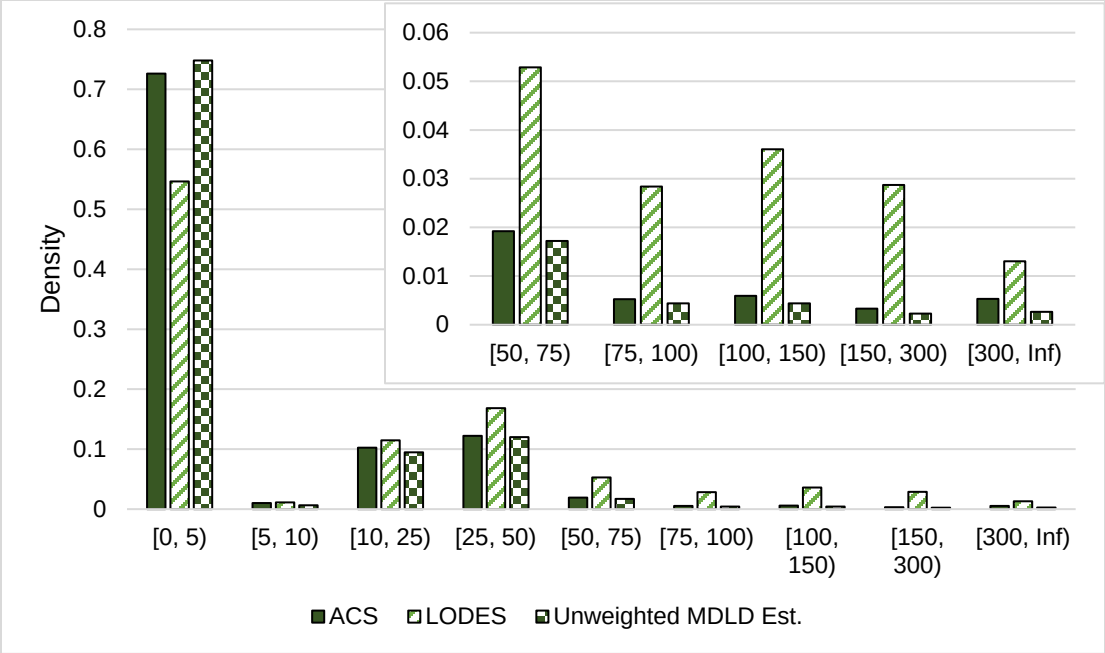


Figure 15. County-level commuting distance distribution

In summary, the validation results demonstrate the reliable performance of the proposed home and work location identification algorithms.

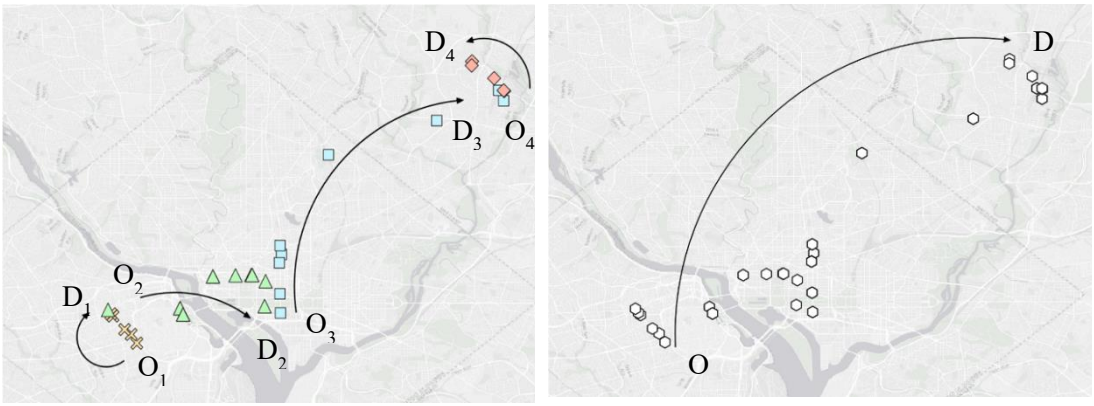
4.2. Tour and Trip Identification

Trips are the unit of analysis for almost all transportation applications. Traditional data sources, such as travel surveys, record the details of trip information. The mobile device location datasets, on the other hand, do not directly provide trip information. Location sightings can be continuously recorded while a device moves, stops, or stays static. However, these changes in status are not recorded in the raw MDLD. As a result, researchers must rely on trip identification algorithms to extract trip information from raw data.

While the literature review and practice scan reveal many methods to identify trips, a key issue complicates the trip identification process and affects the accuracy and credibility of the algorithms, which is ignorance of the difference between linked and unlinked trips. Existing trip identification methods can only identify unlinked trips but not linked trips. For instance, a single transit commute trip with longer than five minutes of waiting at the origin and transfer transit stations would be identified as three unlinked trips with existing methods: (1) a walking trip from home to the origin transit station; (2) a transit trip from the origin transit station to the transfer station; and (3) another transit trip from the transfer station to the final destination. However, for the purpose of tracking individual mobility behavior, the tour and linked trip notions would provide additional useful information to enhance monitoring of the mobility behavior of individuals. Additionally, being able to determine the tour and linked trip information provides a great opportunity to compare the statistic derived from MDLD with traditional travel surveys more accurately. Also, the tour information can be utilized to improve the current travel mode imputation algorithms using MDLD data. It should also be noted that the tour-based approach is necessary to identify the true origins and destinations of long-distance trips.

Figure 16 illustrates how the proposed tour-based algorithm can be used to link the trips together. The four unlinked trips from Figure 16 (a), i.e., a driving trip from home to the metro station (O1 to D1), the first leg of a metro trip to the transfer point (O2 to D2), the second leg of the metro trip on another metro line (O3 to D3), and a walking trip to the work location (O4 to D4), form one linked trip from home to work

in Figure 16 (b). The linked trip from home to work and an additional linked trip from work to home construct one complete home-based-work tour in this case.



(a). Multiple Unlinked Person Trips

(b). One Linked Person Home-to-Work Trip

Figure 16. Tour identification and trip chaining demonstration

4.2.1. Home-based Tour Identification

The algorithm requires devices' identified home locations as input. The home-based tour identification processes a device's locations every day, from 4 a.m.-3:59 a.m. the next day, or "trip day". All sightings between two at-home observations will be considered as a home-based tour. As long-distance trips demonstrate distinct spatio-temporal characteristics compared to short-distance trips, the tours are classified based on their distance feature. Long-distance tours are defined as tours in which a device is observed equal to or more than 50 miles away from its home location. To be consistent with the common practice in travel surveys, the device starts and ends the trip day at home. In the next step, the sightings of each device are separated into two groups: sightings on short-distance tours and sightings on long-distance tours. Finally,

short-distance tours go through a daily short-distance trip identification and long-distance tours go through a monthly long-distance trip identification.

4.2.2. Trip Identification for Short-distance Tours

It is possible that some sightings do not belong to any trips (i.e. stationary points). For each sighting within the same tour, a recursive algorithm based on the decision tree model is utilized to identify if the sighting is stationary or moving. The decision tree considers six attributes, i.e. the great circle distance, time interval, and speed between the current sighting and the previous and next sightings. The decision tree has three hyper-parameters: a distance threshold of 300 meters, a time threshold of 5 minutes, and a speed threshold of 3 miles per hour. The speed threshold is used to identify if a sighting is recorded on the move, and the distance and time thresholds are used to identify trip ends.

The recursive algorithm checks every sighting to identify if they start a new trip or belong to the same trip as the previous sighting (Figure 17). If the previous sighting is not on a trip (i.e. a stationary sighting) the current sighting starts a trip if it has a speed faster than 3 mph to the next sighting. If the previous sighting is on a trip, the following rules are checked to identify if the current sighting belongs to the same trip, stops the trip, or starts a new trip:

- If a sighting has a speed greater than 3 mph from the previous sighting, the sighting belongs to the same trip as its previous sighting.

- If a sighting has a speed slower than 3 mph from the previous sighting and is more than 300 meters away from the previous sightings, the sighting does not belong to the same trip as its previous sighting. If the speed to the next sighting is also slower than 3mph, the current sighting simply terminates the trip; otherwise, it becomes the start of a new trip.
- If a sighting has a speed slower than 3 mph from the previous sighting and is within 300 meters from the previous sighting, the cumulative dwell time for all the consecutive sightings meeting such criteria is computed and checked: 1) if the cumulative dwell time is less than five minutes, the current sighting belongs to the same trip, 2) otherwise, it terminates the trip if the speed to the sighting is slower than 3 mph or starts a new trip if the speed to the next sighting is faster than 3 mph.

The algorithm may identify a local movement as a trip if the device moves within a stay location. To filter out such trips, all trips shorter than 300 meters are removed as a post-processing step.

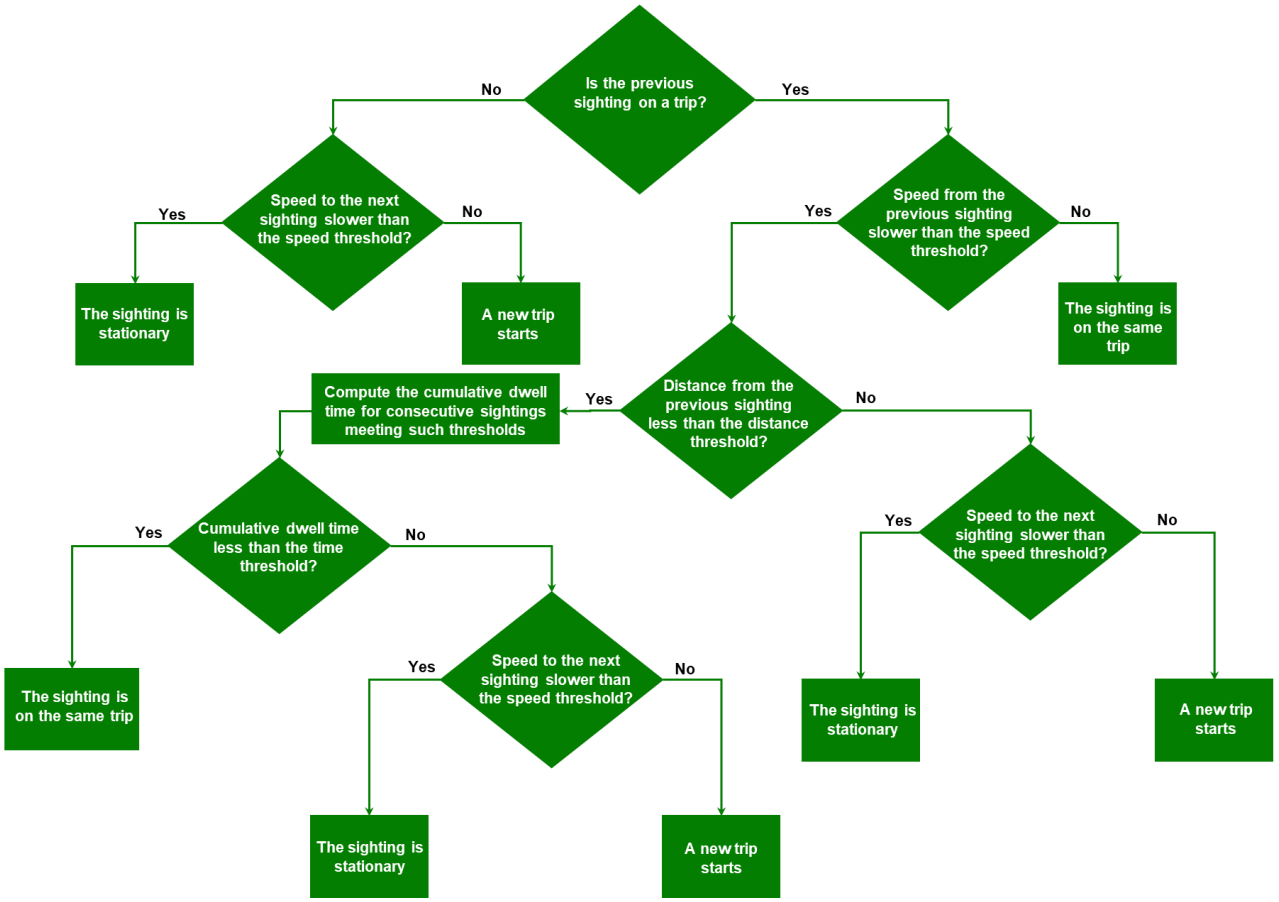


Figure 17. Recursive trip identification algorithm for short-distance tours

4.2.3. Trip Identification for Long-distance Tours

Trip identification for long-distance tours follows a different procedure due to the different nature of long-distance trips. To start, all device sightings on long-distance tours for the entire month are filtered.

4.2.3.1. Stop and primary destination identification

A recursive trip identification, similar to that described in section 4.2.2, is applied, but with a larger time threshold of 30 minutes instead of 5 minutes, meaning that a trip ends only if the device stays in a location for more than 30 minutes. In this step, all the trip ends are identified and named as “secondary stops”. Primary stops are then defined from the secondary stops. Primary stops on a long-distance tour are places where the device stays for a significant amount of time and/or from which the device makes local trips. In order to identify the primary stops, each secondary stop is checked against the following criteria:

- The duration of stay in the secondary stop is longer than two hours and during the stay, the device exits and reenters the secondary stop
- The duration of stay at a location is longer than 24 hours
- The secondary stop is the home location

Furthermore, the primary destination of a tour is defined as the farthest stop that is located at least 50 miles away from the home location of the device. The primary destination is unique in each long-distance tour and is identified from the primary

stops. If no primary stop fulfills the requirement, the primary destination is then identified from the secondary stops.

4.2.3.2. Subtour identification

A subtour is considered a segment of a long-distance tour that falls between two primary stops. Therefore, all sightings between two primary stops are considered to be on the same subtour.

4.2.3.3. Trip extraction

If a long-distance tour does not have a primary destination or has the same primary destination as the identified work location, the short-distance trip identification algorithm (with a time threshold of five minutes) is applied to all the sightings in the tour. If a tour has a primary destination different from the fixed work location, the long-distance trip identification algorithm with a time threshold of 30 minutes is applied to sightings between two different primary stops, and the short-distance trip identification recursive algorithm with a time threshold of 5 minutes is applied to sightings around the same primary stop (local trips around a primary stop on a long-distance tour).

Finally, all the tour, subtour, and trip information are consolidated to provide a complete travel diary of a device.

The complete framework of long-distance trip identification is presented in Figure 18.

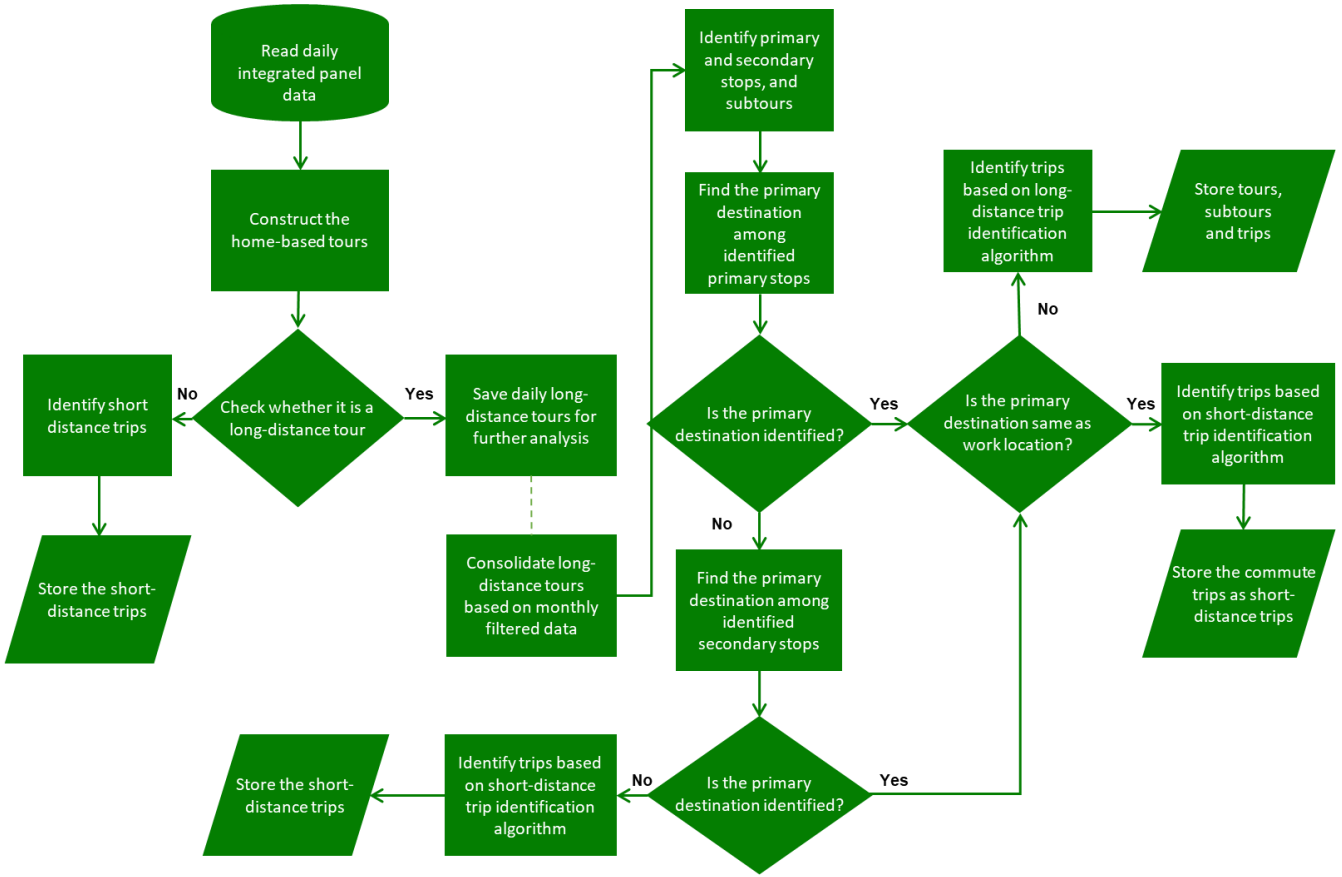


Figure 18. Trip identification framework for long-distance tour

4.3. Trip Mode Detection

Following the trip identification algorithm, a framework has been proposed to impute the travel mode of the trips based on the characteristics of the trip (143). The major contribution of this proposed algorithm is to combine the advantages of a single-layer model and deep neural network to accurately detect the travel mode of the trips.

4.3.1. Data Collection for Travel Mode Imputation

A ground truth dataset with true labels is required to train the proposed supervised learning algorithm. This study used smartphone GPS survey data collected from 300 Washington D.C. urban travelers through a smartphone application that records trips for each survey subject. The survey app functions are illustrated in Figure 19:

- GPS location tracking: the app automatically records users' location information. The frequency of recording was automatically adjusted based on whether the user was moving or static in order to save battery consumption. Typically, the time interval between two location records was 30 seconds when users were moving and between 10 to 30 minutes when users were static depending on the battery status.
- Opt-in trip information survey: the app periodically popped up survey questions to record trip purposes and the travel modes for the users' recorded trips. This information was verified by a follow-up travel diary survey and used as the ground-truth travel mode dataset with labels to train the mode detection model.

- Data uploading: for the sake of battery and cellular data usage, the app did not automatically upload data to the online database unless the device was plugged in and connected to a Wi-Fi network. Alternatively, the user could manually upload survey records by pressing the button “Press to Upload”

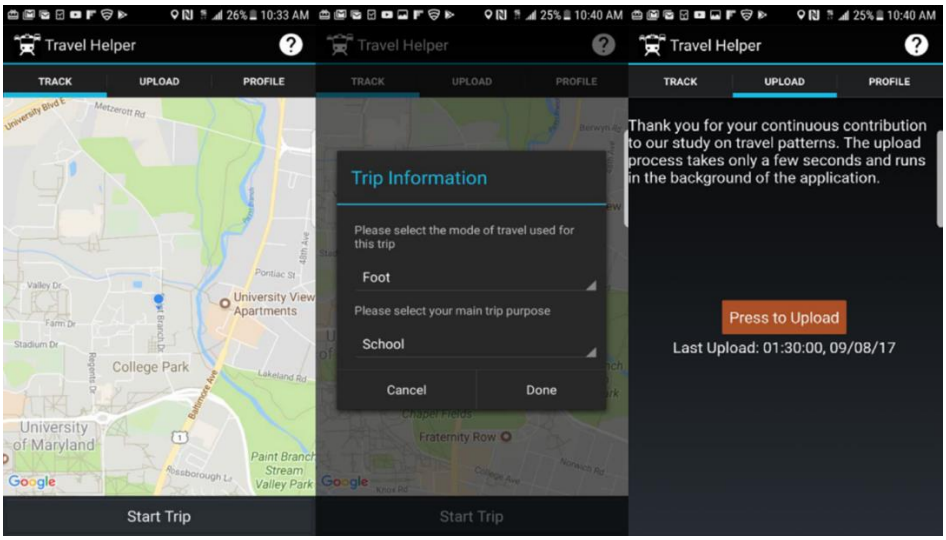


Figure 19. The user interface of the smartphone GPS data survey app

A total of 1009 validated trips were specified with travel mode information. Of these 1009 trips, 19.3% were auto trips 15.9% were bus trips, 52.9% were metro or rail trips, and 11.9% trips were walk/bike trips. Since the survey was targeted toward urbanized areas, a higher percentage of metro and bus trips were captured. This additional bus and rail evidence helps to enhance the understanding of their characteristics and improve the goodness-of-fit of the model for those travel modes.

4.3.2. Construction of Classification Features

Table 5 summarizes the trajectory features that are considered in this study. These features are selected to differentiate the modes as much as possible, For instance, the average speed can be used to distinguish walk mode from other modes. The maximum speed further helps differentiate walk trips from auto or bus trips that encounter severe traffic congestion making their average speed close to non-motorized trips. The overall data recording frequency can be utilized to identify metro trips as other travel modes typically do not suffer from significant GPS disruptions.

Table 5. Trajectory features description

Variables	Descriptions
Trip distance	The trip distance is computed as the sum of the distances between two successive location points in this trip
Trip time	The difference between the timestamps of the trip start and the trip end.
OD Euclidean distance	The shortest Euclidean distance between the origin and destination of the trip
Average speed	The average speed is calculated as the trip distance divided by the trip time
Max. instantaneous speed	The maximum value in the set of instantaneous speeds directly collected by the smartphone app during the trip.
Speed quantiles	The 5^{th}, 25^{th}, 50^{th}, 75^{th}, 95^{th} percentiles of speed are also calculated for each trip.
Average data record	The number of data points recorded during the trip divided by the trip time.

In addition to these features, this study used the available metro, rail, and bus networks to construct additional features (Figure 20). In specific, the average distances to transportation networks were added as geographic features. From a

location point in a trip trajectory, the nearest metro and rail line was first identified using the network shown in Figure 20.

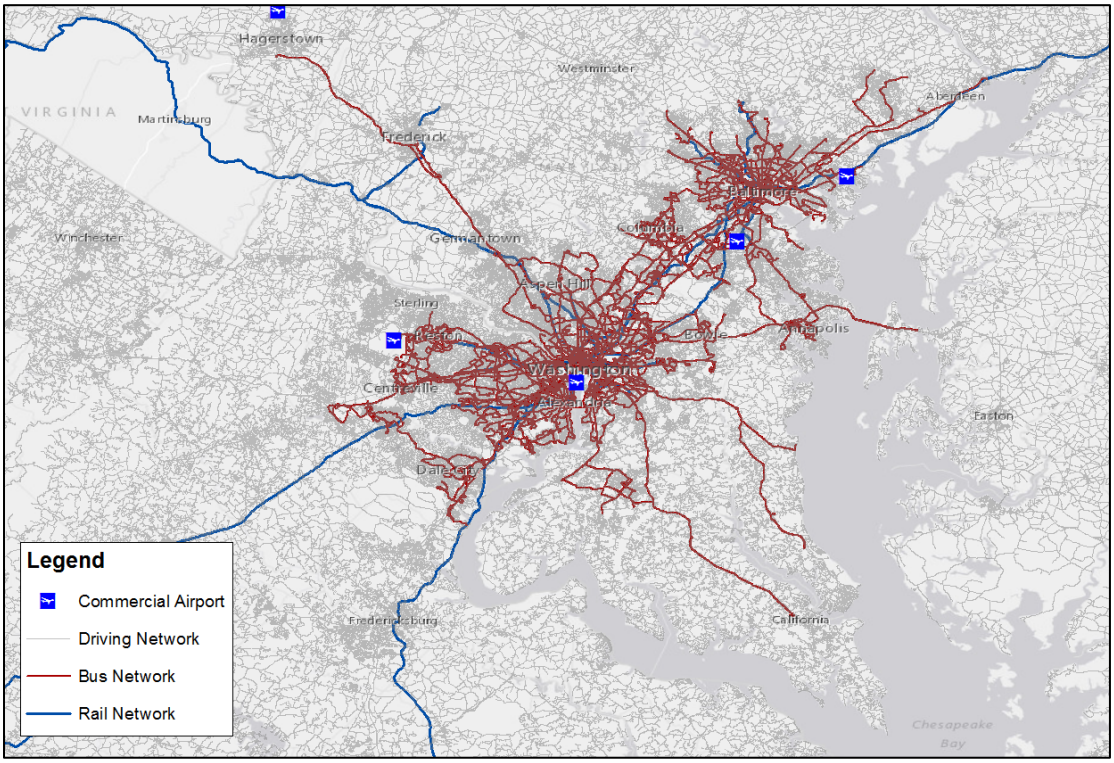


Figure 20. Multimodal transportation network of the study area

The shortest Euclidean distance for each trajectory location point in the trip is calculated then. These distances are then averaged to measure the average adjacency of the trip to the metro and rail systems. Similarly, the average distance to the nearest bus line network was calculated which is deemed essential in improving the accuracy of the mode detection. To comprehensively assess the network effect, the rail network was extracted from the National Transportation Atlas Database (NTAD). The General Transit Feed Specification (GTFS) bus shapefiles have also been collected from 31

regional and local agencies and bus services to construct the bus network. The predictive power of adding these network features is assessed in Section 4.3.4.

4.3.3. Model Structure

This study proposes a mode detection algorithm based on a wide and deep learning approach as illustrated in Figure 21 (143).

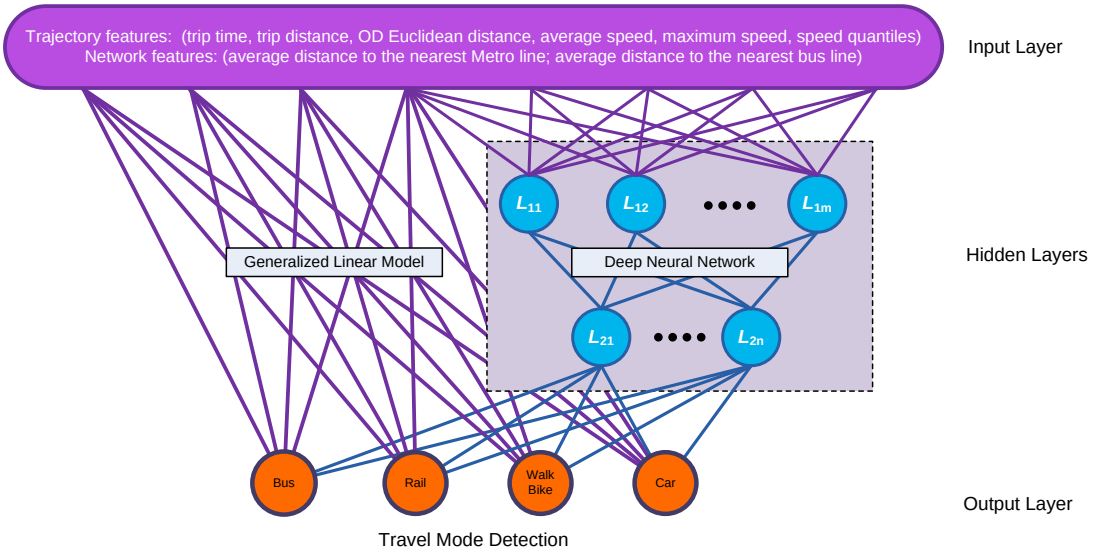


Figure 21. The wide and deep learning framework

A generalized linear model and a deep neural network are jointly trained based on the features constructed using the passively collected data gathered from the survey. Because of the structure of the model, the model is capable of generalizing rules and memorizing specific exceptions at the same time which leads to a superior prediction accuracy compared to stand-alone generalized linear models, and stand-alone deep neural network (DNN) models. To further examine the performance of the proposed

model, benchmark ensemble models and Random Forest have also been trained for comparison purposes. All models were trained and fine-tuned using the TensorFlow platform in Python.

Both trajectory features and network features are used in the Wide and Deep model. These features are all continuous and were normalized to the range of [0,1]. Two hidden layers in the DNN are illustrated in Figure 21 with m neurons and n neurons, respectively. The number of layers and the number of neurons in each layer can be fine-tuned. In the empirical test of this study, three hidden layers have been used and different numbers of neurons were also tested.

Denoting y as the label for travel mode, x as the vector of prediction features, beta as the vector of model parameters, and b as the unobservable heterogeneity, the wide component of the model is formulated as a generalized linear model. In this case, a multinomial logit model is considered:

$$Pr(Y = y) = \frac{\exp\left(\boldsymbol{\beta}_y^T \mathbf{x}_y + b_y\right)}{\sum_i \exp\left(\boldsymbol{\beta}_y^T \mathbf{x}_i + b_i\right)} \quad (2)$$

Where Y is the prediction, x_y is a vector of d features for mode y, $\boldsymbol{\beta}$ is a d-dimensional vector of model parameters, and b is the bias. Then a three-layer DNN has been specified as the deep component. The variables were fed into the hidden layers of the DNN to perform the following computation in each hidden layer (144).

$$a^{(l+1)} = f\left(\gamma^{(l)} \cdot a^{(l)} + b^{(l)}\right) \quad (3)$$

Where a, γ, and b denote the activations, DNN parameters, and heterogeneity at the l--th layer respectively. f denotes the activation function, which defines the output of the neuron node given an input. RELU (rectified linear units) has been used as the activation function, f(z) = max (0, z). In practice, the RELU function works robust and has a better computational efficiency in comparison with the other activation functions (144) although it is not differentiable when z = 0. The combination of the generalized linear model and the DNN represents a model of wide and deep learning that can be jointly trained using the weighted sum of the log-odds as the objective function. The prediction function for the wide and deep learning model is:

$$Pr(Y = y) = \sigma\left(\boldsymbol{\beta}_y^T \mathbf{x}_y + \gamma^{(l_f)} \cdot a^{(l_f)} + b\right) \quad (4)$$

where *Pr* denotes the prediction of the joint model, $\boldsymbol{\beta}_y^T$ denotes the vector of parameters for the linear model component, and $\gamma^{(l_f)}$ denotes the finalized parameters on the final activations of the DNN component, labeled as $a^{(l_f)}$. $\sigma(\cdot)$ is the sigmoid function.

Back-propagation of the gradients was employed to jointly train the model. Gradients were defined from the mode detection to the generalized linear model and the DNN hidden layers based on the weighted sum of the log-odds from both models (144). A number of optimization algorithms were tested to reach the optimal level of training loss and reasonable training time at the same time, including AdaGrad (145), RMSProp (146), and Adam Optimization (147). RMSProp seems to yield the highest

goodness of fit with acceptable computational efficiency. The models reported in this study were trained within 20~60 seconds on a regular Macintosh machine.

AdaGrad algorithm employs adaptive learning rates with a decay factor (145). The rates can adapt to different gradients, which makes the algorithm suitable for high-dimensional problems. However, the descent of AdaGrad can be too fast and the algorithm can get trapped in a local optimum. RMSProp and Adam algorithms address the issue by introducing an exponential decay of past gradients, so that the most recent gradient will have a higher influence on the gradient used in the current iteration. These adaptive optimization algorithms are all tested in this research to compare their performance on the mode detection application.

Finally, with a Random Forest model or a Wide and Deep model trained, a 10-fold cross-validation was conducted to test the performance. To ensure randomness and reasonable stability of the results, a subset of the dataset was randomly sampled using 10 random seeds, and then each subset was partitioned into ten equal-sized subsamples. In each fold of the 10-fold validation, one subsample was retained as the hold-out test sample, and the model was trained using the remaining nine subsamples.

4.3.4. Empirical Results

Several state-of-practice and state-of-art algorithms including ensemble models (AdaBoost and Bagging have been tested, Bagging is reported in this section because of its better performance), Random Forest, generalized linear model, and wide and deep neural model (various optimizers have been tested, with AdaGrad and RMSProp

reported) were trained and compared using the collected dataset. The prediction accuracy of 10-fold cross-validation has been used to measure the performance of the candidate models. For each round of the validation, 10 random seeds were used to ensure the stability of the validation results. Grid search and random search have been used to fine-tune the hyper-parameters in the candidate models.

Table 6 summarizes the performance measures of the models. The first finding is that the addition of multimodal network features has significantly boosted the model performance. Both the ensemble model and Random Forest have shown improved model prediction accuracy after the inclusion of network features. From the 10-fold cross-validation with 10 random seeds, the Random Forest model can get 89.6% of the travel modes in the testing data accurately detected. Also, the benchmark Random Forest model outperforms the Generalized Linear model, suggesting that rule-based generalization using features such as the maximum speed or the distance to nearby transit stations could play a significant role in travel mode detection.

Table 6. Goodness of fit measures for different travel mode detection models

Model	Total Loss	Average Loss	Average Accuracy
Generalized Linear Model	26.0	0.299	0.867
Ensemble (Bagging, without network features)	104.4	1.060	0.755
Ensemble (Bagging, with network features)	84.0	0.860	0.804
Random Forest (RF, without network features)	52.2	0.600	0.808
Random Forest (RF, with network features)	17.4	0.193	0.894
Wide and Deep Model (AdaGrad Optimizer, with network features)	6.7	0.076	0.957
Wide and Deep Model (RMSProp Optimizer, without network features)	17.2	0.197	0.921
Wide and Deep Model (RMSProp Optimizer, with network features)	4.0	0.045	0.976

The wide and deep model combines the advantages of the DNN and the Generalized Linear Model, and can boost the prediction accuracy to above 95%. With 400 neuron nodes coded in the first hidden layer and a default optimizer, AdaGrad, the average prediction accuracy of the model reaches 95.7%. Equivalently, the reduction of prediction errors achieved by using a joint Wide and Deep model is more than 50%. The best Wide and Deep model with RMSProp optimizer can reach 97.6% prediction accuracy. A deeper look at the confusion matrices (Table 7) offers more insights into the performance of the model. The sums of rows and columns may differ due to the random sees used.

In total, a comparison of four models, RF and Wide and Deep with and without network features was conducted. From the confusion matrix, the prediction accuracy for each mode can be evaluated separately. For instance, the first row of Table 7

suggests that 195 car trips were reported in the testing dataset while 135 of them were classified correctly by the RF model without network features.

Table 7. Confusion matrix comparison of RF model and the wide and deep learning model

RF without network features		**10-Fold Cross-Validation: Detected Travel Mode**				
		Car	Metro	Bus	Walk	**Recall:**
Reported Travel Mode	Car	135	34	23	3	69.2%
	Metro	23	479	25	7	89.7%
	Bus	23	42	90	5	56.3%
	Non-motorized	1	4	4	111	92.5%
Precision:		74.2%	85.7%	63.4%	88.1%	80.8%
RF with network features		**10-Fold Cross-Validation: Detected Travel Mode**				
		Car	Metro	Bus	Walk	**Recall:**
Reported Travel Mode	Car	181	4	7	3	92.8%
	Metro	7	507	13	7	95.0%
	Bus	15	43	101	1	63.1%
	Non-motorized	0	5	2	113	94.2%
Precision:		89.2%	90.7%	82.1%	91.1%	89.4%
Wide-Deep, without network features		**10-Fold Cross-Validation: Detected Travel Mode**				
		Car	Metro	Bus	Walk	**Recall:**
Reported Travel Mode	Car	172	8	13	2	88.2%
	Metro	8	508	16	2	95.1%
	Bus	11	14	132	3	82.5%
	Non-motorized	0	2	1	117	97.5%
Precision:		90.1%	95.5%	81.5%	94.4%	92.1%
Wide-Deep, with network features		**10-Fold Cross-Validation: Detected Travel Mode**				
		Car	Metro	Bus	Walk	**Recall:**
Reported Travel Mode	Car	194	1	0	0	99.5%
	Metro	0	525	8	1	98.3%
	Bus	1	10	149	0	93.1%
	Non-motorized	1	1	1	117	97.5%
Precision:		99.0%	97.8%	94.3%	99.2%	97.6%

By adding the network features, the precision and recall accuracies were significantly increased. Overall, one of the benchmark models, Random Forest with network features, did a decent job in detecting car, Metro, and non-motorized modes. However, the precision of detecting bus mode still falls short. Comparing the Random Forest with the Wide and Deep model, it is clear that the latter did extremely well in the detection of Metro and bus trips. Even without the network features, the Wide and Deep model can get to a similar level of accuracy to the RF model with the network features. The Wide-Deep model without network features achieves a precision accuracy of 82.5% for the bus mode, compared to 56.3% in the RF model. By adding the network features to the Wide-Deep model, the precision/recall accuracies rocket to above 93%. It is worth noting that this study only conducted a standard grid search in combination with optimizers. By researching the fine-tuning of the joint model, the accuracy could be further improved. This could direct the path of future studies.

Chapter 5: MDLD in Action for Pandemic Studies

Since the first case of the novel coronavirus disease (COVID-19) was confirmed in Wuhan, China, social distancing has been promoted worldwide, including in the United States, as a major community mitigation strategy. However, our understanding remains limited in how people would react to such control measures, as well as how people would resume their normal behaviors when those orders were relaxed. This dissertation proposes a framework to quantify the impact of COVID-19 on mobility and provide insights to analyze human mobility behavior throughout the pandemic (136, 148).

5.1. Methodology

After cleaning the data, identifying the home and work locations, and extracting the trip information, based on the methodologies described in chapters 3 and 4, this study investigated the mobility behavior of communities throughout the COVID-19 pandemic.

To fully leverage the near real-time mobility insights from the MDLD, two additional methodological steps were needed to be introduced. First, a weighting method that can convert the sample movements observed in the MDLD to population-level statistics. Next, introducing an index that could summarize different aspects of communities' mobility patterns into a single metric that could capture the impact of COVID-19 on mobility.

5.1.1. Weighting

In spite of MDLD's high penetration rate among the population, statistics derived from the MDLD still need to be weighted to represent population-level statistics. The devices available in the dataset are a sample of all individuals in the population, so it is necessary to consider device-level weights. In addition to the device-level weights, MDLD might only capture a sample of all trips conducted by the individuals in the data. Therefore, trip-level weights are also needed.

As the goal of this study was to provide near real-time mobility statistics updates, a simple county-level device weighting has been applied to obtain weights for devices. To derive device-level weights, the home county for each device has been specified based on the identified home location. The weight for each device was calculated based on the number of devices observed in the device's imputed home county divided by the population of the county, so all devices residing in a county would have the same device-level weights. For instance, if the sample includes 100 devices in a county with a population of 2,000, each device would be assigned a weight of 20. The population of each county has been obtained from the U.S. Census Bureau.

For the trip level weights, the number of trips per person (trip rate) has been calculated for each state during an average weekday in the first two weeks of February 2020 from the sample with the assumption that the February travel behavior was not impacted by the COVID-19 pandemic. Then the trip rate number has also been calculated for each state from the most recent national household travel survey, 2017 NHTS. Then a state-level trip rate has been calculated by dividing the NHTS

trip rate by the observed trip rate during the pre-pandemic period. These weights are used for the entire study period.

5.1.2. Core Mobility Metrics

After completing the extraction of population-level trips from MDLD, all information was summarized into several core mobility metrics that are critical for a better understanding of the national mobility pattern before and during the pandemic. Table 8 shows the list of metrics calculated at the county, state, and national levels.

Table 8. List of core mobility metrics calculated to capture the COVID-19 impact on mobility

Current Metrics	Description
% staying home	Percentage of residents staying at home (i.e., no trips more than one mile away from home)
trips/person	Average number of trips taken per person.
% out-of-county trips	The percent of all trips taken that travel out of a county.
% out-of-state trips	The percent of all trips taken that travel out of a state.
miles traveled/person	Average person-miles traveled on all modes per person per day (car, train, bus, plane, bike, walk, etc.)
#work trips/person	Number of daily work trips per person (where a "work trip" is defined as going to or coming home from work)
#non-work trips/person	Number of daily non-work trips per person. (e.g. grocery, restaurant, park, etc.).

5.1.3. Social Distancing Index

In addition to calculating the core mobility metrics, this dissertation explored the construction of a single index that could capture the mobility changes and portray individual efforts in social distancing by considering the various measurements of human mobility.

To properly design the structure of the Social Distancing Index (SDI), the existing indices from various fields have been reviewed. There are two main types of indices: category-based indices and score-based ones. The category-based indices explain the proposed objective by categories. For example, the Pandemic Severity Index (PSI) classified the case fatality ratio (CFR) of disease into five categories (from one to

five) (149), and the Modified Mercalli Intensity Scale evaluates the severity of an earthquake by categorizing it into twelve levels from I to XII (150). On the other hand, score-based indices usually define a score from zero to one hundred to differentiate objectives and rank them in order. For example, the US. News State ranking creates a score that covers eight topics on people's needs in each state and assigns different weights to those topics based on the survey data (151). Bloomberg Global Health Index is another score-based index that ranks countries in terms of healthiness by giving them a rate between zero and one hundred (152). In short, category-based indices are usually built upon a single variable and the score-based ones are more capable of integrating multiple metrics to be more informative.

In this effort, SDI was designed as a score-based index, which gives a 0-100 score to each geographical area, e.g. a state or county, and measures to what extent area residents and visitors practice social distancing in terms of mobility aspects. Zero indicates no social distancing and one hundred indicates perfect social distancing compared with the benchmark days before the COVID-19 outbreak. The benchmark values for the core metrics are computed using data from the weekdays (Monday to Friday) during the first two weeks of February. Thereafter, the changes in people's mobility patterns are captured by the percentage reduction of the corresponding metrics in Table 9 (noted as X_2,…, X_5) as input. The absolute changes in the percentage of residents staying home (noted as X_1) also serve as input. The percentage reductions are absolute values between 0 and 100%. Any increase is standardized as 0% in the calculation.

Table 9. Descriptive statistics for the core metrics

Index	Metric	Min	Max	Mean	Median
1	% staying home	13.0	58.0	26.1 SD: 7.6	25.0
2	#work trips/person	0.14	1.49	0.48 SD: 0.18	0.46
3	#non-work trips/person	1.39	3.90	2.64 SD: 0.37	2.65
4	miles traveled/person	15.6	113.4	52.3 SD: 14.3	52.1
5	Out-of-county trips (in thousands)	7	28845	5339 SD: 5299	3597

By jointly considering the travel behaviors of region residents and visitors, the equation for computing SDI is given as follows:

$$SDI = [(\beta_1 X_1 + 0.01 \times (100 - X_1) \times (\beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4)] \times (1 - \beta_5) + \beta_5 X_5 \quad (5)$$

Where $\beta_1 = 1$ and $\beta_2 + \beta_3 + \beta_4 = 1$.

The first part of the equation focuses on resident level and the second part on out-of-county trips. β_5 is thus the weight assigned to behavior changes regarding out-of-county trips. For the resident trips, we use the percentage of residents staying home to account for residents who do not make trips longer than 1 mile from home, so the weight is simply one ($\beta_1 = 1$). For people not staying home (travelers), the percentage of which is 100-X_1, I use a weighted sum of percentage reductions in the number of work and non-work trips made daily and the average distance traveled per

person. When individuals make more work and non-work trips, and travel longer distances, they are considered to practice less social distancing. The weights for each variable should sum up to one ($\beta_2 + \beta_3 + \beta_4 = 1$) so that resident travelers are comparable to residents staying at home.

To assign appropriate weights to each variable, both actual observations and conceptual guidelines are consulted. Firstly, the relative ratio between resident trips and out-of-county trips nationwide is about four to one. Hence, a weight of 0.2 was assigned to β_5. Secondly, it is widely observed that people have significantly reduced travel distances so the index should not give the large percentage reduction in distances traveled the same weight as the reduction in the number of trips. Meanwhile, the reductions in the number of trips are more informative with regards to people's reaction to the stay-at-home mandates. Thus, the reduction in the number of trips is considered twice as important as that in distance traveled and a weight of 0.3 was assigned to β_4. Moreover, as suggested by government agencies, people are highly encouraged to reduce non-essential trips. Therefore, the index should be designed to factor in the reduction in non-essential trips, which is estimated twice as important as the reduction in essential trips. Work trips are intuitively considered essential trips and non-work trips could include both essential and non-essential. Based on the 2017 National Household Travel Survey (NHTS) Travel Profile (153), the traveler ratio between essential and non-essential non-work trips is approximately 1:2. Therefore, the relative ratio between the percentage reduction of work and non-work trips is 1:1.67. According to the constraint $\beta_2 + \beta_3 + \beta_4 = 1$, 0.25 and 0.45 were assigned to β_2 and β_3 accordingly. The SDI is eventually computed as follows:

$$SDI = [(X_1 + 0.01 \times (100 - X_1) \times (0.25X_2 + 0.45X_3 + 0.3X_4)] \times 0.8 + 0.2X_6 \quad (6)$$

It should be noted that the weights are partially determined by certain assumptions. For example, the reduction of trips is considered more important than the reduction of travel distances when measuring the social distancing strength. The sensitivity of SDI scores was evaluated as the relative weights between the trip and distance reduction estimates changed. It was observed that assigning a higher weight to the distance reduction estimates (β_4) lead to larger absolute values and standard deviations of SDI scores. When $\beta_4 = 1$, the largest absolute values and standard deviations of SDI scores are observed. Although the magnitude of SDI scores has changed, both spatial and temporal trends stayed the same in general. Therefore, such changes in weight assignments shall not yield inconsistent inferences when comparing the social distancing practices between different regions and periods.

5.2. Results

To add more context to the observed mobility changes during the COVID-19 outbreak, the mobility metrics are integrated with COVID-19 case data (154).

5.2.1. The effectiveness of the Social Distancing Index (SDI)

The effectiveness and reasonableness of the proposed SDI were examined by reviewing its temporal change from February 2, 2020, to May 30, 2020, and the spatial variation by states for the entire nation (Figure 22).

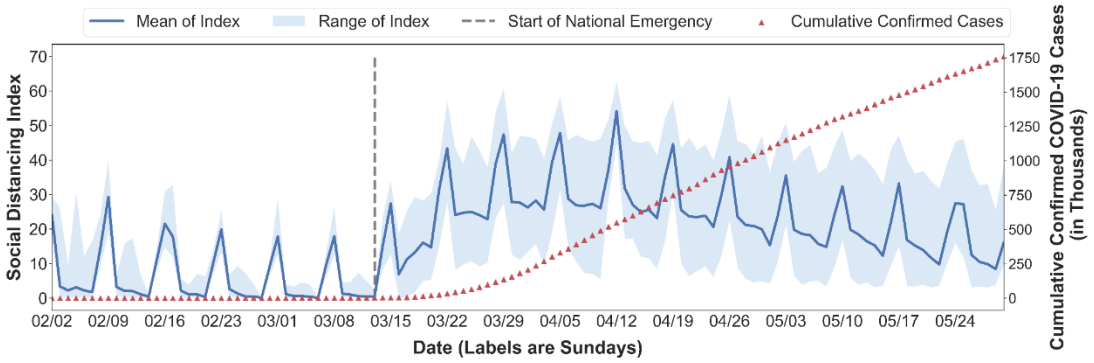


Figure 22. Temporal changes of state-level Social Distancing Index

The proposed SDI is sensitive to people's behavior changes and is capable of reflecting the mobility changes accordingly. The SDI changes clearly indicate that people stay home more and travel less on weekends, especially on Sundays, and people traveled less on Memorial Day (May 25, 2020) compared with a normal Monday. During the study period, people practiced significantly more social distancing nationwide after President Trump declared a national emergency concerning the COVID-19 outbreak. The national emergency declaration immediately triggered people's responses on weekdays beginning March 16 and on weekends of the following weeks: March 22, March 29, and April 5. In addition, the range of the index became wider after March 16, indicating that people from different states were having distinct responses to the national emergency announcement.

After the week of March 23, a general plateau was observed in terms of social distancing practices. Beginning April 6, there was a tendency toward less social distancing in some states. One week later, a similar trend appeared across the entire nation. The possible reasons are twofold. First, people became less attentive to the

outbreak as the outbreak persisted at the time. Moreover, because of the widespread economic impacts of the pandemic, some people could no longer afford to maintain social distancing. As people reduce social distance measures, there was no significant slowdown in the number of reported COVID-19 cases.

5.2.2. State-level Mobility Pattern Changes

Following the national emergency declaration, the mandatory stay-at-home orders issued by most states triggered a second wave of strengthened social distancing. This influence of government mandates on human behavior can also be seen when some states began reopening: states that chose to lift stay-at-home mandates early saw an acceleration in social distancing relaxation. The SDI is computed for all states for thirteen consecutive weeks from March 1 to May 30, 2020, in Figure 23. Five stages are defined based on the general trend from all states: pre-pandemic (before March 13), behavior change (March 13 to March 22), government orders and holding steady (March 23 to April 12), quarantine fatigue (April 13 to April 26), and partial reopening and stay-at-home order lifting (April 27 till the end of the study period).

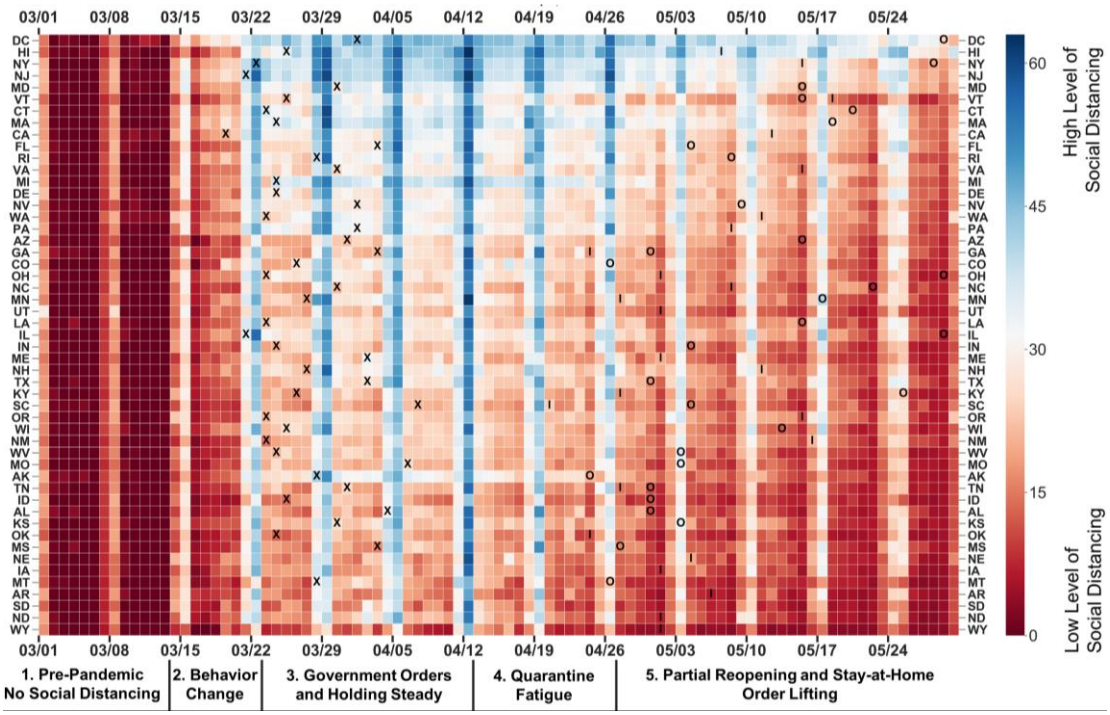


Figure 23. Social Distancing Index heatmap for all states

Figure 23 shows the level of SDI scores for all states during the study period. Each pixel in the graph indicates the level of social distancing for one specific state on a specific day, where blue stands for more social distancing practiced and red for less. The “X” marker indicates the start date of state-wide stay-at-home orders. The “O” marker indicated the order lifting date. The “I” marker indicates the start date of state-wide partial reopening if different from the order lifting date. The states are sorted in descending order by their SDI scores on the last weekday (May 29, 2020). The top five regions that were performing more social distancing are the District of Columbia, Hawaii, New York, New Jersey, and Maryland, all of which issued stay-at-home orders. Meanwhile, the states practicing less social distancing are Wyoming, North Dakota, South Dakota, Arkansas, and Montana, most of which did not issue stay-at-

home mandates. One other consideration is that on the East and West Coasts, it is possible that people practice more social distancing because they were exposed to the infection risk for a longer period and were aware of higher infection risk with higher population density.

In Figure 24, the cumulative number of confirmed cases on May 30 2020 for the top five and bottom five states were examined. After the stay-at-home orders were issued, all 10 states experienced an increase in SDI, but the bottom five states generally had lower scores of SDIs. This implies that the local severity of the COVID-19 outbreak played a significant role in people's decision-making. Although all ten states experienced a decrease in SDI after April 13, a sharp decline was observed following the partial re-opening and/or stay-at-home order lifting in New York, Massachusetts, and Alaska. This implied that people in those states were willing to maintain more social distancing for a longer period, but the early reopening discouraged social distancing behavior. The influence of early reopening in Alaska appeared after two weeks when the increase in confirmed cases accelerated. Similar impacts of reopening can be observed in California, Montana, Oregon, and West Virginia, where the low level of SDI and increasing trend of confirmed cases raised concerns about a second local outbreak.

In Figure 24, the blue dots stand for SDI scores on weekdays and the orange dots for SDI scores on weekends. The red triangular dots stand for the daily cumulative number of confirmed COVID-19 cases. The grey line stands for the start date of the

state stay-at-home order. The green line marks the stay-at-home order lifting date and the green dashed line marks the date of state partial reopening.

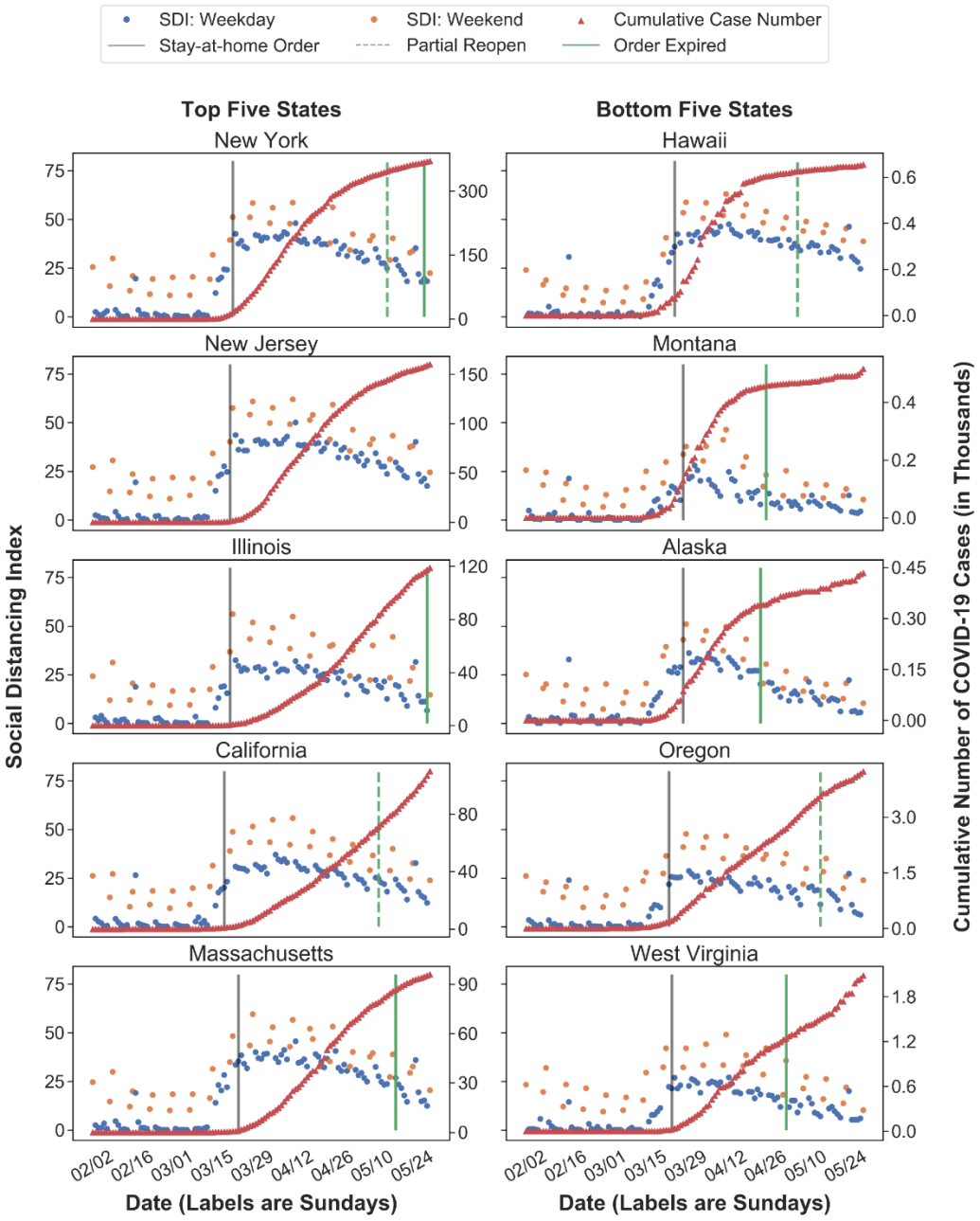


Figure 24. Temporal changes of Social Distancing Index in the top five and bottom five states regarding the cumulative number of confirmed cases.

The Spearman's rank correlation coefficient between the infection rates and the SDI scores for those ten states has also been evaluated for the entire study period. Table 10 summarizes the results. Since the SDI scores on weekends are systematically higher than those on weekdays, only the weekdays' observations were used to compute the correlation coefficients.

Table 10. Spearman's rank correlation coefficient between SDI and infection rate for the top five and bottom five states regarding the cumulative number of confirmed cases.

Top five states	Infection Rate		Bottom five states	Infection Rate	
	Cumulative	New		Cumulative	New
New York	0.658	0.663	Hawaii	0.744	0.713
New Jersey	0.689	0.669	Montana	0.611	0.604
Illinois	0.573	0.582	Alaska	0.660	0.661
California	0.594	0.599	Oregon	0.619	0.594
Massachusetts	0.614	0.619	West Virginia	0.651	0.643

The cumulative infection rate is defined as the cumulative number of confirmed COVID-19 cases per thousand population, and the new infection rate as the number of new confirmed cases daily per thousand population.

In Table 10, a stronger correlation was observed between SDI and new infection rate than that between SDI and cumulative infection rate, suggesting that people were paying close attention to the outbreak development and have been practicing less social distancing. The stronger correlation between SDI and new infection rates in Hawaii, New Jersey, Massachusetts, and New York implies that people in those states were more attentive during the pandemic compared to other states. Those states also

have a flatter curve of the cumulative number of confirmed cases at the end of the study period.

5.2.3. County-level Mobility Pattern Changes

SDI is also informative at the county level. Figure 25 demonstrates the temporal changes of SDI for the top ten counties with regard to the cumulative number of confirmed cases on May 30, 2020. The counties in New York performed strict social distancing, which helped "flatten the curve" of cumulative confirmed cases. The high levels of SDI in Middlesex County, MA, Wayne County, MI, and Hudson County, NJ have also slowed down the outbreak. However, a relaxation of social distancing was observed after the partial reopening and the expiration of stay-at-home orders. In the meantime, Los Angeles County, CA, and Philadelphia County, PA were among regions that needed to strengthen their social distancing practices as their SDI scores were lower than other counties in similar circumstances and their confirmed cases showed an increasing trend at a rapid pace.

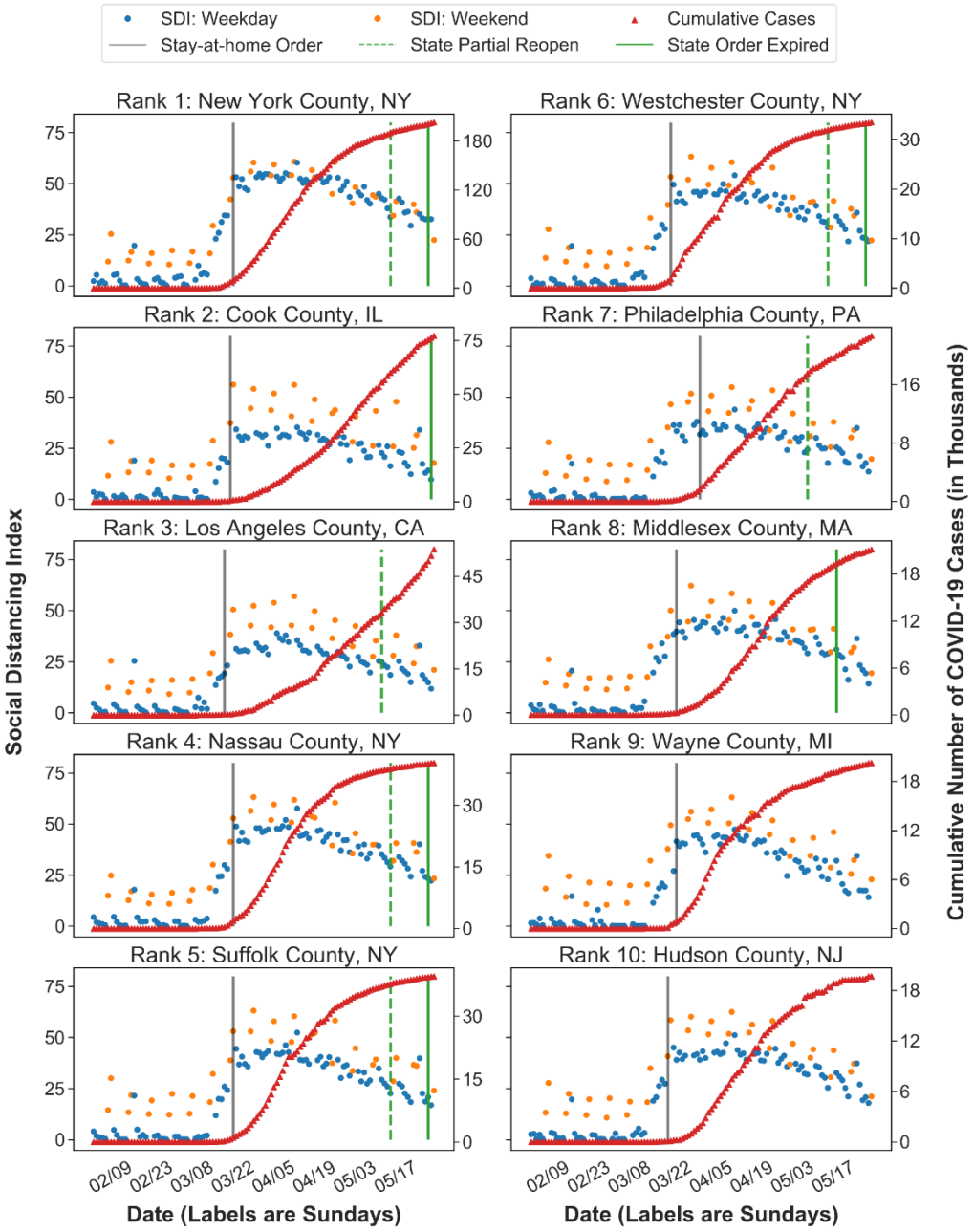


Figure 25. Temporal changes of Social Distancing Index in the top ten counties according to the cumulative number of confirmed cases.

The correlation between the infection rates and the SDI scores was also evaluated for the top ten counties with regard to the cumulative number of confirmed cases. Table

11 summarizes the results. In general, stronger correlations between the infection rates and the SDI scores were observed in the counties with higher SDI scores. Moreover, the counties with smaller correlation coefficients between SDI and new infection rates tended to have an increasing trend in the cumulative number of confirmed cases at the end of the study period.

Table 11. Spearman's rank correlation between SDI and infection rate for the top ten counties regarding the cumulative number of confirmed cases

Top ten counties	Infection Rate		Top ten counties	Infection Rate	
	Cumulative	New		Cumulative	New
New York County, NY	0.734	0.746	Westchester County, NY	0.709	0.721
Cook County, IL	0.590	0.608	Philadelphia County, PA	0.695	0.655
Los Angeles County, CA	0.636	0.651	Middlesex County, MA	0.708	0.705
Nassau County, NY	0.706	0.715	Wayne County, MI	0.698	0.679
Suffolk County, NY	0.689	0.670	Hudson County, NJ	0.730	0.732

5.3. Summary and Discussion

During the COVID-19 pandemic, data-driven tools that can proved insight into human mobility behavior have been of paramount importance. This dissertation introduced the real-world observation of human movements from MDLD, to study the impact of non-pharmaceutical interventions. By studying the travel behaviors of people across the United States, a score-based Social Distancing Index (SDI) was developed to capture people's actual social distancing behaviors. Monitoring the SDI patterns, both spatially and temporally, enables policymakers to evaluate the effectiveness of related policies and to involve data-informed decision-making for

public health. In addition, SDI boosts public and community awareness regarding the ongoing situation for where they are living. People can use insights from SDI to evaluate the potential risks in their neighborhoods.

Being exploratory research, this study could be further improved in several directions. Firstly, the basic mobility metrics could be generated considering regional differences. Specifically, the current definition of the stay-at-home population may introduce some bias due to different individual behaviors between residents in rural and urban areas. For example, many people living in rural regions still must make long trips to shop for essential goods while people in urban areas have a higher chance of obtaining essential items nearby (within 1 mile from home) and thus are more likely to be identified as staying at home. Secondly, adding more mobility metrics to the SDI could contribute to the comprehensiveness of the index. For instance, the trip purposes could be inferred by integrating MDLD and point of interest (POI) data. Identifying where people visit could provide the opportunity to distinguish between essential and non-essential trips, in addition to distinguishing between work and non-work trips. Thirdly, variables measuring the relationship between human movements and disease transmission could be extremely valuable. Although it may be difficult to retrieve details such as contact tracing information from MDLD, the aggregated measurements can also be significant indicators, such as trips from and to the heavily infected areas that yield potential exposure and disease transmission in the study, on top of out-of-county trips that are currently included. Moreover, an expert survey on improving the weight assignments to different variables in SDI may also contribute to better construction of the index.

Another future research direction is to integrate SDI with existing epidemiological frameworks, such as compartment models. A variable of interest in these frameworks is to understand how the input variables evolve during the course of the outbreak. Certain policies, such as mobility restrictions, can significantly reduce certain input variables like the reproduction factor of the disease. SDI can be employed in these models to enhance the input prediction in compartmental models.

Chapter 6: MDLD in Action for Disaster Evacuation

Understanding individuals' behavior during natural disasters is of paramount importance for the local, state, and federal government agencies hoping to be prepared for these extreme situations. In this study, a novel framework is introduced to construct evacuation patterns and analyze individuals' decisions (155). Hurricane Irma and the state of Florida have been selected as the case study for implementing the framework and testing the results.

6.1. Introduction

In September 2017, Hurricane Irma prompted officials to issue one of the largest evacuation orders in U.S. history. Over six million people were ordered to evacuate their residences due to Irma's landfall in Florida, Georgia, and South Carolina. Mandatory and voluntary evacuation orders were issued before the landfall of the storm, on both the Atlantic and Gulf coasts. 84 deaths were reported just in the state of Florida due to either direct effects of Hurricane Irma such as drowning or indirect causes such as vehicle accidents during the evacuation. The immense scale of hurricanes and the dependence of the evacuation management on how people behave during these disasters highlighted the importance of studying the evacuation patterns of the people in such situations.

6.2. Data

6.2.1. Location Data

The primary dataset used in this study is the MDLD of anonymized devices from LBS data sources. Based on meteorological history, Irma developed from a tropical wave near Cape Verde on August 30 and quickly intensified into a category 3 hurricane by August 31 due to the climate condition. On September 4, the storm kept intensifying, making it a Category 5 hurricane.

Therefore, based on the timeline of Hurricane Irma's evolution, the month of August 2017 is chosen to identify the home location of the users within the state of Florida with the assumption that users' behavior had not been impacted by the news of Hurricane Irma yet. For the analysis of the mobility behavior and to understand the evacuation pattern of the residents in Florida, the data from the entire month of September 2017 were analyzed.

6.2.2. Evacuation Zone Data

In addition to the location data, gathering information regarding evacuation order evolution was necessary to understand the individuals' behavior. The Florida Division of Emergency Management provided the spatial polygon of evacuation zones for the counties with defined evacuation zones. However, for the information regarding evacuation orders by county and zones, no single source provided comprehensive details. The webpage of Florida governor, Rick Scott, had one of the

most complete information regarding the issuance of evacuation orders as of 9/9/2017. However, several counties, particularly in the north of Florida, issued evacuation orders on 9/10/2017. Also, many counties upgraded evacuation orders from voluntary to mandatory on or after 9/9/2017. Therefore, data from several sources has been compiled to provide a complete picture of the evacuation orders. The final Florida map by evacuation order and date during Hurricane Irma is shown in Figure 26 (156). Besides the evacuation map, open-source parcel-level information for the entire state of Florida was obtained. The data were gathered by the Florida Department of Revenue, County Property Appraisers, and the University of Florida GeoPlan Center. This layer contains residential home type information that has been used in the parameter selection process for the home location identification algorithm.

Also, to measure the impact of living in low-lying residences on the evacuation decision, the elevation information was obtained from the digital elevation model (DEM) provided by the University of Florida GeoPlan Center for the entire state of Florida.

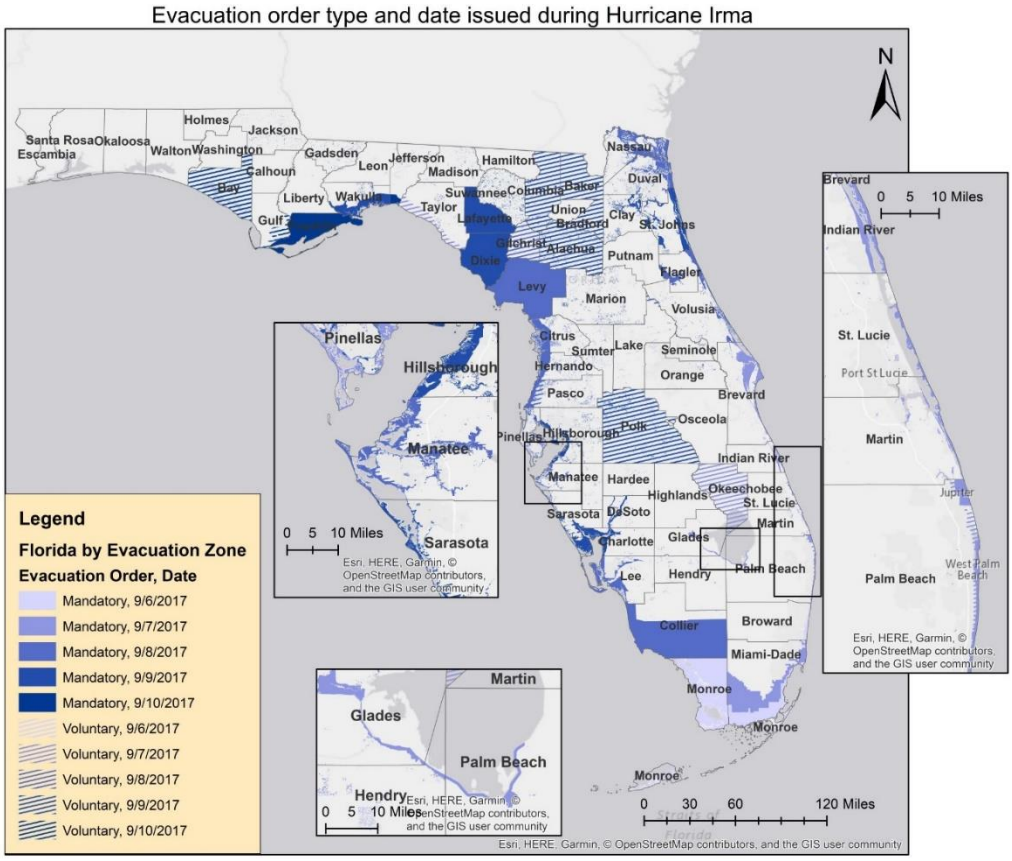


Figure 26. Florida map by evacuation order and date during Hurricane Irma

6.2.3. Socio-Demographic Data

The socio-demographic information such as income, age, and race information was gathered for statistical modeling purposes. To collect this information at the census tract level, 2017 American Community Survey (ACS) 5-year estimates conducted by the United States Census Bureau have been used.

6.3. Methodology

To construct the evacuation behavior pattern, three main steps are designed. The first step is to identify the home location of all devices. Next, a framework is proposed to determine devices that evacuated and to construct their evacuation behavior. Lastly, mobility metrics of devices are calculated to examine the relationship between the evacuation decision and the mobility behavior of the individuals. Figure 27 illustrates the framework structure.

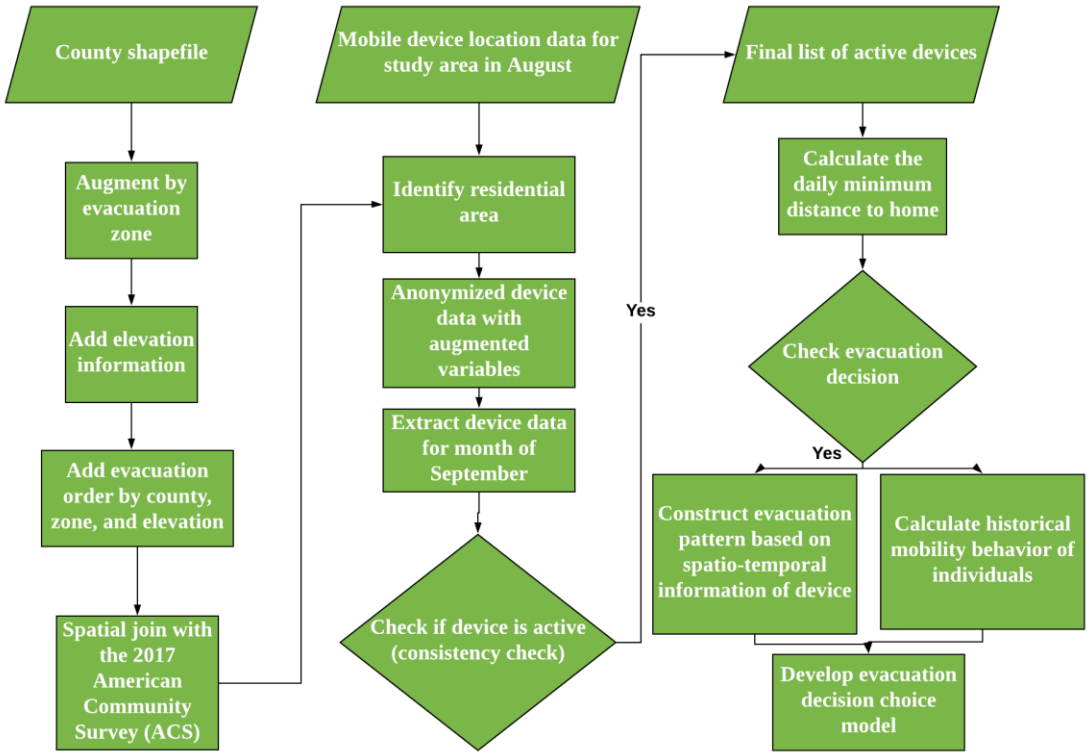


Figure 27. Disaster evacuation analysis framework flowchart

6.3.1. Home Location Identification

For this application, as the scope of the study was limited to devices within the state of Florida only for one month, a more computationally intensive home location identification algorithm has been developed.

To cluster the sightings of each device, the Density-based spatial clustering of applications with noise (DBSCAN) clustering approach was used. DBSCAN is a clustering algorithm relying on a density-based notion of clusters, designed to discover clusters of sightings regardless of their shapes (72). In addition to a more computationally intensive method, a longer nighttime window from 7 pm to 7 am was considered for the home location identification. Among all the identified clusters determined by algorithm, the home location was defined as the center of the cluster with the highest dwell time and the highest frequency observation, respectively.

6.3.2. Evacuation Detection

After filtering the devices with the inferred home located within the state of Florida, the sighting data of these devices for the entire month of September were extracted to study the evacuation pattern of the residents of Florida during Hurricane Irma.

First, to ensure the persistency and accuracy of the identified home location in August, only devices that have been observed at least once in their August home location during the month of September were kept for further analysis. This check removes devices without any information in September, along with devices that have changed their home location or were observed in Florida during August on a trip.

Next, the identified August home location of each device was intersected with the augmented shapefile to specify the corresponding county, evacuation zone, elevation information, and socio-demographic attributes of each device. The census-tract level socio-demographic attributes were added to all devices that resided in the census tract.

The next step was to define evacuation based on the observed trajectories for each device. An evacuation identification method was developed based on the distance of the users' sightings to their inferred home location during the landfall of Hurricane Irma. For this purpose, the daily minimum distance between the device's sightings and their identified August home location was calculated for each device for the entire month of September. A 1-mile threshold was selected as the evacuation criterion to determine whether each individual evacuated. If individuals were not observed within a 1-mile radius of their home locations within the hurricane study period, they were considered as individuals who evacuated their home location. The former Florida Governor, Rick Scott, declared a state of emergency on September 4, and within the next six days, 57 of the 67 counties issued evacuation orders. Eventually, Hurricane Irma made landfall on Cudjoe Key on September 10 as a category 4 hurricane and exited Florida into Georgia on September 11, after being significantly weakened. Thus, the period between September 4 and September 12 was chosen as the hurricane study period for determining the evacuation decision of the individuals.

6.3.3. Historical Mobility Behavior Pattern

In addition to constructing the evacuation pattern, in this dissertation, the relationship between individuals' mobility behavior before the disaster and its impact on their evacuation decisions have been investigated. In particular, two important mobility aspects of the individuals, the number of trips and convex hull set information of each individual have been calculated daily for the entire month of August. The convex hull is defined as the smallest convex set that contains all the spatial sightings. Convex hull has been widely used for understanding human mobility behavior based on location trajectories in the literature (39, 157).

6.4. Constructing the Evacuation Pattern

In addition to the evacuation decision, departure and reentry dates are of paramount importance in disaster evacuation management. Therefore, the minimum daily distance to home measure has been used to investigate the distribution of the departure and reentry dates. For the individuals who evacuated, the latest day before the evacuation in which they were seen in the 1-mile radius of their identified home was chosen as their departure date. Similarly, the earliest day after the evacuation, in which they were seen within the 1-mile radius of their identified home was selected as their re-entry date. Estimating the departure and reentry date provides the opportunity to further investigate the relationships between departure dates and other influential factors such as the evacuation order date.

Destination choice is another important decision component. While an increase in short-distance evacuations increases the demand for sheltering resources, it reduces the stress on the transportation network as well as the overall cost of the evacuation operation. In this study, the maximum of the minimum daily distances from the inferred home location was used as a proxy for the evacuation destination. Also, the impact of living in a low-lying residential area on individuals' evacuation decisions was empirically examined by controlling for the type of evacuation order received.

6.4.1. Stay or Evacuate

By implementing the home location identification algorithm discussed in section 6.3.1 on more than 6 billion observations for the devices that were observed in Florida during August, the home location of 1,050,472 devices was identified. Among this set of devices, 1,002,858 devices resided within the state of Florida. Extracting the information of these devices for September, 5,677,549,347 sightings were filtered from the MDLD data for further investigations. The persistency checks were conducted to remove inactive devices during September as well as eliminate devices that did not have any sightings in the vicinity of their identified home location. The final list of devices includes 807,623 active devices. The minimum distance from the identified home location was calculated daily for all users. Then the proposed framework for evacuation identification was employed to determine the evacuation decision, departure and reentry dates of the evacuees. A summary of the rate of evacuation by each evacuation order type is shown in Table 12.

Table 12. Evacuation decision based on the evacuation order received

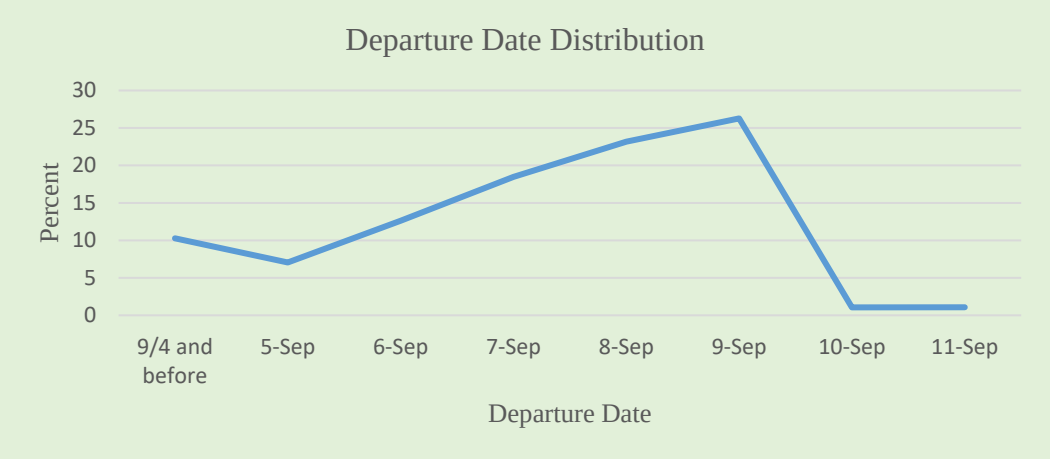
	No Evacuation Order		Voluntary Evacuation Order		Mandatory Evacuation Order		Entire State	
	Number	Ratio	Number	Ratio	Number	Ratio	Number	Ratio
Evacuated	187285	32.98	38524	33.68	72628	57.92	298437	36.9
Not Evacuated	380547	67.02	75868	66.32	52771	42.08	509186	63.1
Total	567832	100	114392	100	125399	100	807623	100

Based on the results summarized in Table 12, 57.92% of the individuals who received mandatory evacuation orders evacuated their homes while this ratio was considerably lower for people who received voluntary evacuation or no evacuation order (33.68% and 32.98%, respectively). These results are in accordance with the results of a telephone poll conducted on October 17, 2017, that showed 57% of people followed the mandatory evacuation order and in general, 33% of Floridians evacuated their homes (158).

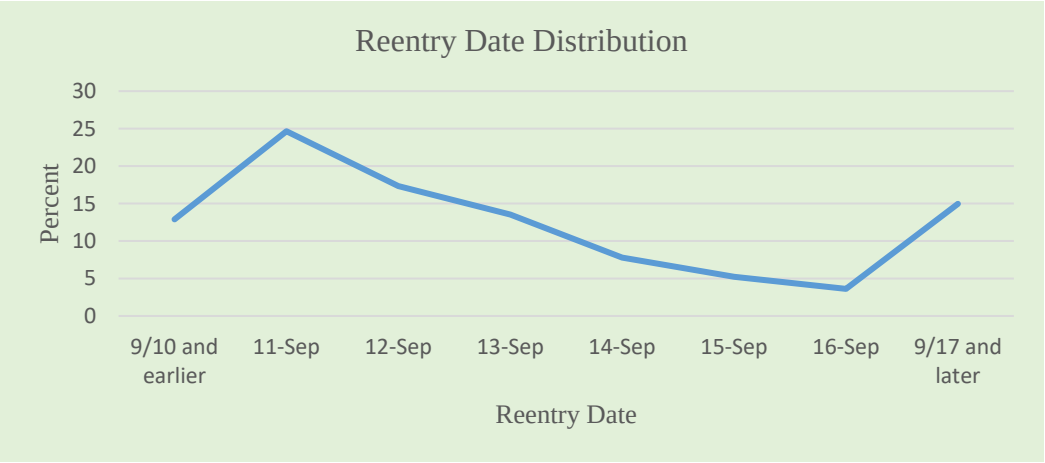
6.4.2. Departure and Reentry Date Distribution

Departure and reentry date choices are becoming increasingly important for emergency and transportation practitioners as well as state and government agencies. I tried to estimate the departure and reentry date distribution by employing the method discussed in section 6.4. It should be acknowledged that this approach might lead to some inaccuracy in capturing the actual departure and reentry dates for devices that may have lost their connections to the network either due to power outages or losing cell network services during and after the hurricane landfall. However, comparing the results with the conducted survey for the same region show a consistent pattern (120). A summary of the results is presented in Figure 28.

Based on the results, the majority of the evacuations occurred from September 8 to September 9, with September 9 being the peak with 26.27%. Although the majority of evacuations happened in the last three days before Irma's landfall, the results showed that a considerable number of individuals evacuated their homes 5 days or earlier in advance, with 7.04% of people evacuated on September 5 and 10.28% evacuated before September 5. This high rate of early evacuation might be due to the fact that some counties started to issue evacuation orders as early as September 5. Increased implementation of time-phased evacuation plans can be another reason for this observation. Finally, only 2.13% of the evacuees left their homes after September 10.



(a) Departure date distribution



(b) Reentry date distribution

Figure 28. Departure and reentry date distribution

On the other hand, reentry date distribution was smoother in comparison to the departure date, with a peak of 24.65% observed on September 11. This was expected since regions do not become livable at once after a disaster. Besides, agencies do not provide returning plans for the impacted areas. Therefore, people usually decide to re-enter their residence in a way that minimizes any impedance such as traffic. Moreover, the results indicated that about 12.89% of the evacuees returned to their homes on September 10 or earlier. This observation has also been observed in a survey as well mainly due to the updates on the hurricane path. Individuals who evacuated earlier may have concluded that their residences were no longer at risk (120).

To delve more into the departure date distribution, the effect of the evacuation order date on the departure date was investigated for all the regions. The majority of the

individuals who received evacuation orders on September 6 departed their homes on September 7 and September 8 while individuals who received evacuation orders on September 7 mostly chose to leave their homes from September 7 to September 9. The same trend can be observed for the people who were ordered to evacuate their homes on September 8. 34.53% of them decided to leave their residences on the following day. As it got closer to the landfall of the hurricane, the impact of the evacuation order date on the individuals' actual departure date decision diminished. The majority of evacuees who were ordered to evacuate on September 9 and September 10 had already left their residences before the receiving of the evacuation order. Figure 29 is color-coded by the evacuation order date.

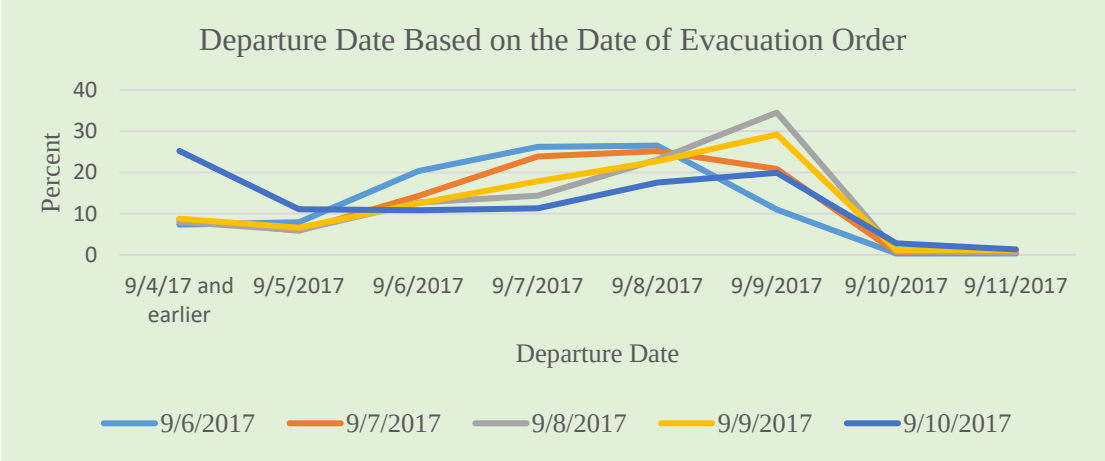


Figure 29. Relationship between departure date and evacuation order date

6.4.3. Destination Choice: Distance to Evacuation Destination

The overall distribution of distance to evacuation destination followed a similar trend among evacuees regardless of the evacuation orders. However, on average, evacuees who received mandatory evacuation orders sought farther locations. The trend is

shown in Figure 30. While about 43% of the evacuees who received voluntary or no evacuation orders decided to choose a destination within a 20-mile radius of their residential locations, 35.47% of evacuees who received mandatory evacuation orders stayed within the 20-mile radius of their home. The distance distribution also suggests that evacuees tend to choose either a close evacuation destination within their neighborhood or travel farther away to reach a location they perceive safe.

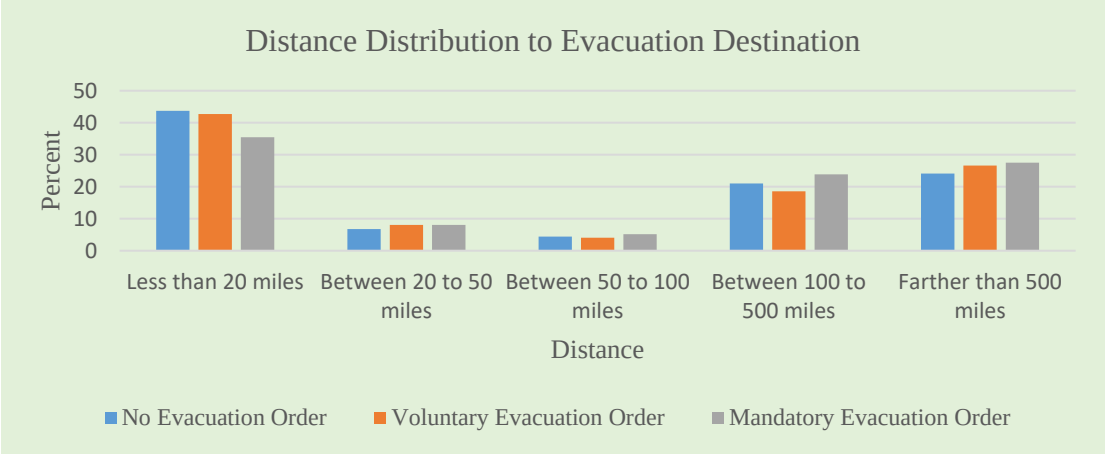


Figure 30. Distribution of evacuation destination distance to the home locations

To dig more into the trend of the evacuation distance, the spatial distribution of the evacuation distance is also illustrated in Figure 31. Evacuees living near the shores tend to travel to farther destinations. This observation is expected as those individuals may perceive a higher risk compared to the people living in the midland.

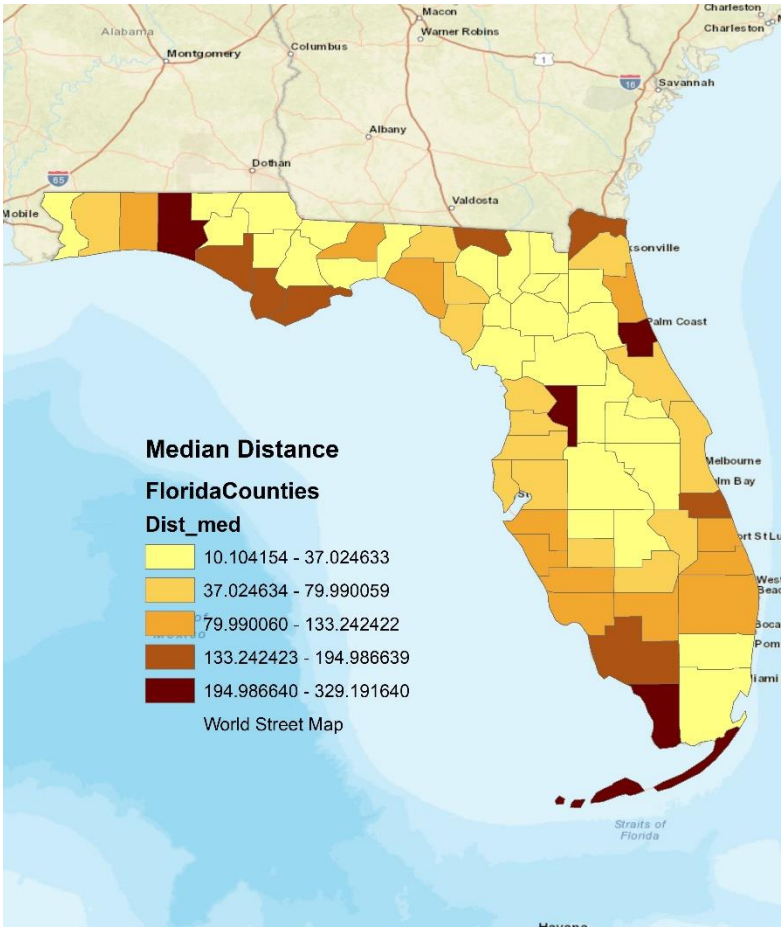


Figure 31. Median distance traveled to evacuation destination at county level

6.4.4. Evacuation Duration Distribution

In terms of evacuation duration, as it is shown in Figure 32, evacuees who received mandatory evacuation orders had a slightly longer evacuation duration. To better understand the spatial trend of the evacuation duration, the average evacuation duration at the county level is also presented in Figure 33. People living in the southern part of Florida had a longer evacuation duration which can be a result of more severe damages to the properties and infrastructures in those specific regions.

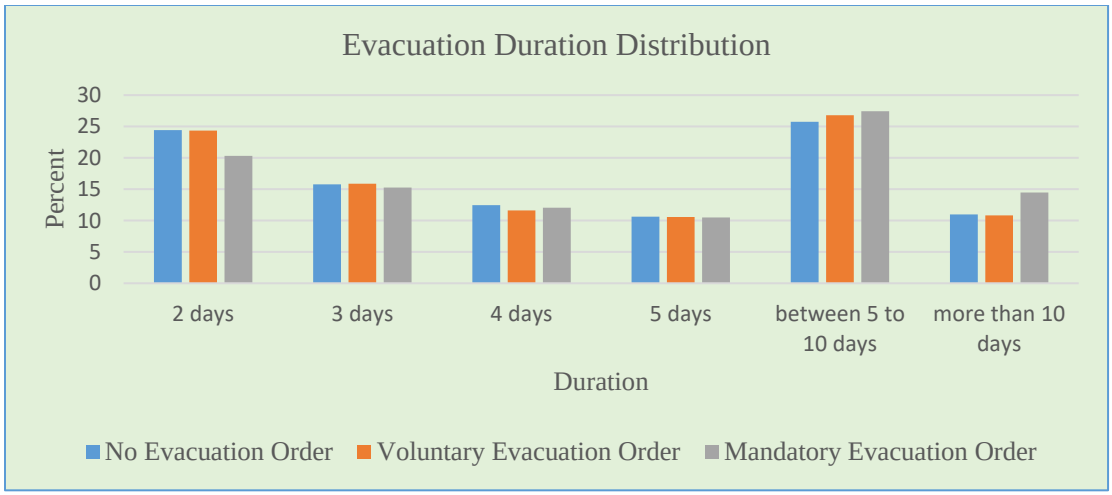


Figure 32. Evacuation duration distribution across different evacuation order groups

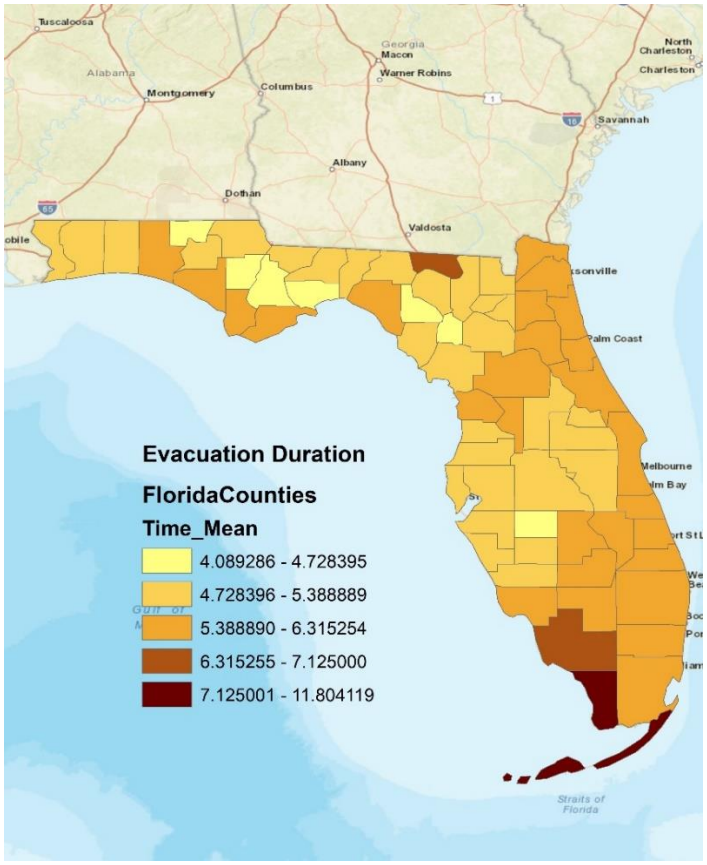


Figure 33. Average evacuation duration at the county level

6.4.5. Impact of Low-Lying Residential Area

The impact of low-lying residential areas on individuals' evacuation decisions has also been investigated. Since there is no strict definition for the low-lying area, three categories were introduced based on the elevation of the residential area; elevation less than 10 meters, between 10 meters to 50 meters, and more than 50 meters. Also to control for the effect of the evacuation orders on individuals' decisions, the evacuation orders were considered. Evacuation rates for each group are presented in Figure 34. It can be seen that the elevation of residential areas has a strong association with people's decision to evacuate. 36.59% of people who had not received any evacuation order but were living in low-lying residential areas decided to leave their homes, while this rate was 28.43% for those in areas with elevation more than 50 meters.

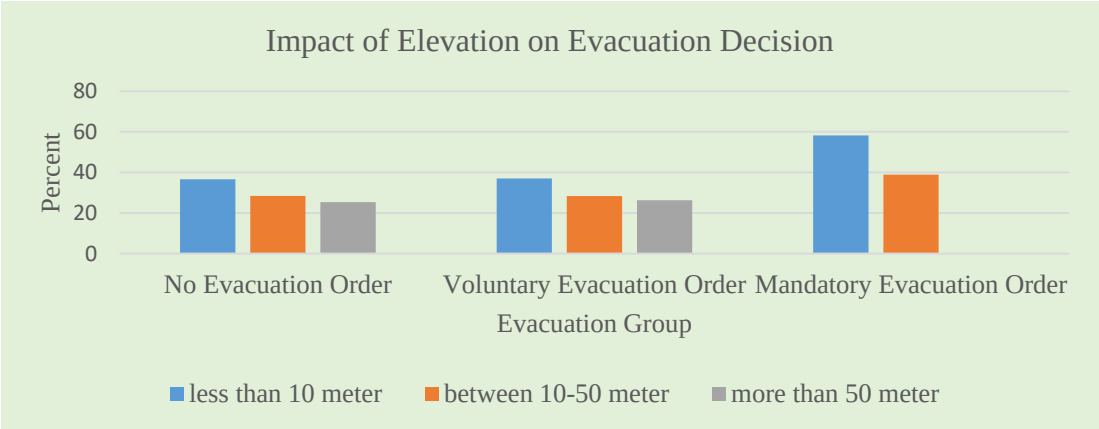


Figure 34. Elevation impacts on evacuation decisions

6.5. Statistical Model

After extracting the evacuation behavior of individuals and comparing the results against existing polls and surveys, this study investigates the statistical linkage between mobility patterns of individuals and their evacuation decisions. The evacuation decision has been well studied in the literature and its importance and implications for agencies have been highlighted. Previous studies revealed the importance of socio-demographic variables such as age, income, and race as well as evacuation orders and the perceived worries and concerns in evacuation decisions. In this dissertation, in addition to those metrics, the importance of individuals' mobility behavior in their decision has also been examined. The individual-level mobility measures including the daily number of trips and the convex hull of each active device were calculated during the entire month of August. The mobility measures were incorporated into the logistic regression model to examine whether those measures are statistically significant in the evacuation decision choice model and whether they can improve the evacuation decision model's accuracy. Table 13 summarizes the list of variables considered for modeling purposes. To develop the statistical model, 3,937 devices were removed from the dataset due to missing socio-demographic attributes.

Table 13. Data description and summary for evacuation choice model

Metric	Definition		Descriptive Statistics			
Categorical Variable			Count		Percentage	
Evacuation Decision	Evacuation decision	0 = did not evacuate,	507605		63.16	
		1 = evacuate	296081		36.84	
Evacuation order	Evacuation order received	0 = none	565178		70.32	
		1 = voluntary,	114038		14.19	
		2 = mandatory	124470		15.49	
Continuous Variable			Min	Median	Max	SD
Elevation	Residential location elevation		-1	6	102	13.86
Median age	Median age of the residential census tract		11.9	41.4	83.3	9.71
Median income	Median income of the residential census tract		8804	54279	2500001	22951
Vehicle availability	Percentage of households with at least one vehicle in the census tract		28.4	96.1	100	5.82
Race - white	Percentage of white population in the census tract		0	0.83	1	0.17
Average number of trip	Average number of trips taken by the individual per day during August		1	5.5	51.4	3.82
Average of convex hull area	Average daily convex hull area of individuals during August		0	48.57	57274.8	510.31

As no evacuation was the base choice in the decision variable, positive coefficients indicate that an increase in variables' value increases the likelihood of evacuation, while a negative sign denotes a decrease in the likelihood of evacuation. The summary of the results is presented in Table 14.

Table 14. Logistic regression models' summary

Variable	Model#1 – logistic model without mobility behavior metrics		Model#2 – logistic model with mobility behavior metrics	
	Estimated coefficient	p-value	Estimated coefficient	p-value
Intercept	3.61E-01	<0.001 ***	4.45 E-01	<0.001 ***
Evacuation order	4.06 E-01	<0.001 ***	4.08 E-01	<0.001 ***
Elevation	-8.60 E-05	<0.001 ***	-8.55 E-05	<0.001 ***
Median age	8.48 E-03	<0.001 ***	8.65 E-03	<0.001 ***
Median income	3.62 E-08	0.766	2.68 E-07	0.028 *
Vehicle availability	-1.57 E-02	<0.001 ***	-1.88 E-02	<0.001 ***
Race - white	2.59 E-01	<0.001 ***	2.44 E-01	<0.001 ***
Average number of trip	-	-	1.03E-02	<0.001 ***
Average of convex hull area	-	-	4.28E-04	<0.001 ***
Number of observation	803686		803686	
Log Likelihood	-516912.5 (df=7)		-513806.2 (df=9)	
AIC	1033839		1027630	
McFadden R2	0.025		0.031	
Models Comparison	P-value (Chi) = <0.001 ***			

As shown in Table 14, two logistic regression models have been developed. Model#1 only includes socio-demographic information, the elevation of residential location, and evacuation order attributes while model#2 utilized mobility behavior metrics in addition to all variables in model#1. In both models, the sign of coefficients for common variables was in line and consistent with previous studies except for the vehicle availability metric. Higher vehicle availability was expected to increase the likelihood of evacuation but in this model, the coefficient was estimated negative. One possible reason for this observation might be due to the low variation of this metric in the study region (the first quantile of vehicle availability was 92.6% and the median was 96.1%). Both mobility metrics turned out to be statistically significant in model#2 and the overall accuracy of the model improved significantly. The estimated

sign of the coefficients was positive which indicates that individuals with more trips per day and a larger mobility footprint are more likely to evacuate their residential location during a disaster.

6.6. Summary and Discussion

The intensity and the frequency of weather-related disasters are expected to increase due to climate change, increase in sea surface temperature, and other related causes (159, 160). In order to be prepared, it is crucial for the state and federal government agencies to understand individuals' behavior before, during, and after a disaster. Most of the research in the literature studied individuals' behavior during these extreme events based on post-disaster surveys. In addition to the small sample size, these surveys are typically prone to several biases, such as observer effect bias and imperfect recall of the evolution of the evacuation process. This dissertation tried to extract information from MDLD to construct several aspects of evacuation patterns by analyzing anonymized individuals' traces.

In this study, the evacuation behavior of 807,623 anonymized individuals was captured by employing the proposed framework on more than 11 billion location sightings. The study results showed that type of evacuation order has a strong impact on individuals' evacuation decisions. Results showed that 57.92% of individuals who received mandatory evacuation orders left their homes while this ratio was 32.98% and 35.68% for smartphone users who received no evacuation order and voluntary evacuation order, respectively.

Irma made its landfall in the mainland U.S. on September 10. The departure date and reentry date analysis conducted in this study demonstrated that the majority of the evacuees left their residences in the last three days leading to the hurricane's landfall, with the peak of evacuation observed on September 9 when 26.27% of evacuees departed their home. However, the returning process was distributed more evenly among days after the landfall. The effect of evacuation orders' dates on individuals' departure date decisions was also empirically examined. It was shown that late evacuation orders (ones that were issued on September 9 and September 10) did not have a strong influence on individuals' departure decisions; while for the regions that received evacuation orders earlier (from September 6 to September 8) an increase was observed in evacuation rate the day after the evacuation order was issued. These findings highlight the importance of issuing evacuation orders at least two days before the hurricane's landfall.

The evacuation distance distribution revealed that the individuals selected to shelter either in the vicinity of their residential area or decided to go to farther away destinations (more than 100 miles away from their home location). It has also been shown that the elevation of residential areas had a strong effect on individuals' evacuation decisions. People living in low-lying regions showed a higher evacuation rate in comparison to people living in mid- and high-elevation regions after controlling for the evacuation order type.

This study also showed that the observed mobility pattern of individuals can play a significant role in improving the accuracy of evacuation decision models. Having

access to historical MDLD provides unique information to the agencies and decision-makers to have a better understanding of the evacuation evolution in their region.

Although analyzing the behavior of smartphone users provides an opportunity to observe the actual behavior of millions of individuals during disasters, several limitations still exist. While the sample size of the MDLD is enormous, it should still be noted that these type of data have their own biases. The other limitation is the fact that post-disaster surveys usually provide a rich set of socio-demographic information and stated preferences of the individuals while MDLD lacks any such information.

Chapter 7: Conclusions and Remarks for Future Work

Understanding people's mobility behavior-i.e., where, when, why, and how people travel is of paramount importance for making decisions and policymaking regarding traffic management and operations, resource allocations, responding to natural disasters, and infrastructure planning. For decades, planners have been relying on two major data sources, i.e. travel survey data and traffic monitoring data (such as roadway traffic volumes, transit ridership information, etc.). The inherent issues and shortcomings of the two data sources, such as small sample size, the cumbersome procedure of obtaining such dataset, and inadequate coverage of travel modes, make the understanding of human mobility patterns costly and prone to known biases.

With the emergence of mobile networks and positioning technologies, mobile device location data have drawn decision-makers and researchers' attention due to their unique potential in analyzing human mobility behavior and understanding travel characteristics. This dissertation constructed a set of frameworks and developed novel algorithms to derive mobility metrics from nationwide MDLD.

The remainder of this chapter begins by summarizing the research contributions and findings of this dissertation, followed by a discussion of the future work directions.

7.1. Summary of Contributions

In chapter 2 first I conducted a comprehensive literature review and practice scan regarding the evolution of mobile device location data and the related advancements in positioning technologies. Then I summarized the research efforts conducted to

extract device- and trip-level information from the MDLD. The literature review is followed by presenting studies that investigated the importance of human mobility behavior in two different study cases, the outbreak of disease and evacuation behavior analysis during natural disasters.

In chapter 3 I introduced the mobile device location data utilized in this study and discussed the data cleaning and preprocessing steps required prior to extracting mobility information from MDLD. The chapter ended by providing a national-level data summary.

Chapter 4 discussed the methodological advancement in inferring device-level and trip-level information from MDLD. A computationally efficient home and work location identification algorithm was introduced in section 4.1. The algorithm was compared with other state-of-practice algorithms and was proven to be both efficient and effective In identifying home and work locations at the national level. In the absence of the individual-level information, the algorithm's outputs were examined against the aggregate level ground truth datasets including ACS estimates and LODES data. Then a novel tour-based trip identification algorithm was introduced to overcome the shortcomings of the existing trip identification algorithms. The tour-based trip identification algorithm leverages the identified home and work location of devices to form tours and enables researchers to differentiate between the long-distance and short-distance tours and link trips together with higher accuracy. The last section of this chapter proposed a new method to impute the travel mode of the trips based on the feature set constructed from both trip trajectory information and the

transportation networks' information of different modes. The empirical results from the proposed algorithm successfully demonstrated its superior performance compared to other state-of-practice and state-of-art algorithms, especially for the modes that are more difficult to be differentiated such as car and bus modes.

In chapter 5, this dissertation developed a framework to quantify the impact of the COVID-19 pandemic on human mobility patterns. The framework was built upon the methodologies described in chapter 4 along with two additional methodological steps (i.e., bi-level weighting and social distancing index construction) to portray a more complete mobility pattern evolution of the communities before and during the pandemic. The national-level, state-level, and county-level mobility pattern trends were investigated to demonstrate the effectiveness and usefulness of such timely data in providing insights to communities and decision-makers.

Chapter 6 extended the human mobility behavior analysis to extreme conditions such as natural disasters. In this chapter, different aspects of evacuation behavior such as evacuation decision, departure time and reentry time, evacuation distance, evacuation duration, and determinant of evacuation decision were studied during a natural disaster. The proposed framework was applied to MDLD for the residents of Florida during the landfall of Hurricane Irma. The proposed framework successfully constructed the evacuation decisions and showed the significance of individuals' historical mobility behavior in their evacuation decisions.

7.2. Future Directions

The applications of MDLD in the transportation domain have grown exponentially since the MDLD data made its debut in the late 1990s. However, there is still room for improvements in the methodologies that are being used to infer human mobility information. I propose the following research directions for future studies:

(1) Preparing an accessible data sandbox with the true device- and trip-level labels with data privacy considerations for the transportation research community. There is a lack of standard and reliable data for transportation researchers to test and develop their algorithms and report consistent accuracy measures for their proposed algorithms. In other fields such as computer science, it is a common practice to use standard datasets to develop and test the performance of different algorithms. This practice has led to significant progress in algorithm development as well as higher transparency in the methodologies.

(2) Human mobility pattern analysis during a pandemic. Chapter 5 only scratches the surface of how insights from human mobility patterns can be used during a pandemic. The core mobility metrics developed for this analysis could be further improved to be tailored toward different communities. For instance, the current definition of the stay-at-home population could be modified in a way to distinguish the inherent differences between the individual mobility behavior in different living environments such as densely urbanized areas versus rural areas. Further research efforts could also be conducted to integrate the mobility measures into the existing epidemiological frameworks such as compartment models as important input variables of the models.

(3) Applications of MDLD in disastrous events. This dissertation shows the feasibility of constructing the evacuation behavior of individuals during a hurricane. Further studies could be conducted to further validate the results of the MDLD and explore the feasibility of providing real-time evacuation information. Improvements in individual-level socio-demographic imputation could also add more context to the MDLD-based outcomes and enables a more in-depth analysis of different evacuation behavior.

(4) Investigating the impact of changes on the mobile device location data streams. The mobile device location data coverage and information collection methods change from time to time due to updates on the privacy protection practices or changes in the technology. A more comprehensive analysis of the impact of these changes should be conducted for a better understanding of the robustness of the derived mobility behavior analysis over time.

Bibliography

1. Gonzalez MC, Hidalgo CA, Barabasi A-L. Understanding individual human mobility patterns. nature. 2008;453(7196):779-82.

2. Xu Y, Shaw S-L, Zhao Z, Yin L, Fang Z, Li Q. Understanding aggregate human mobility patterns using passive mobile phone location data: a home-based approach. Transportation. 2015;42(4):625-46.

3. Levinson D, Kumar A. Activity, travel, and the allocation of time. Journal of the American Planning Association. 1995;61(4):458-70.

4. McNally MG. The four-step model: Emerald Group Publishing Limited; 2007.

5. Wang Z, He SY, Leung Y. Applying mobile phone data to travel behaviour research: A literature review. Travel Behaviour and Society. 2018;11:141-55.

6. Calabrese F, Colonna M, Lovisolo P, Parata D, Ratti C. Real-time urban monitoring using cell phones: A case study in Rome. IEEE transactions on intelligent transportation systems. 2010;12(1):141-51.

7. Alexander L, Jiang S, Murga M, González MC. Origin–destination trips by purpose and time of day inferred from mobile phone data. Transportation research part c: emerging technologies. 2015;58:240-50.

8. Hasan S, Schneider CM, Ukkusuri SV, González MC. Spatiotemporal patterns of urban human mobility. Journal of Statistical Physics. 2013;151(1):304-18.

9. Yabe T, Ukkusuri SV. Effects of income inequality on evacuation, reentry and segregation after disasters. Transportation Research Part D: Transport and Environment. 2020:102260.

10. Frias-Martinez E, Williamson G, Frias-Martinez V, editors. An agent-based model of epidemic spread using human mobility and social network information. 2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing; 2011: IEEE.

11. Batelle. Global Positioning Systems for personal travel surveys: Lexington area travel data collection test. Final Report, Office of Highway Policy Information and Office of Technology Applications, Federal highway Administration, Batelle Transport Division, Columbus.; 1997.

12. Yalamanchili L, Pendyala RM, Prabaharan N, Chakravarthy P. Analysis of global positioning system-based data collection methods for capturing multistop trip-chaining behavior. Transportation Research Record. 1999;1660(1):58-65.

13. Wolf J. Using GPS data loggers to replace travel diaries in the collection of travel data: Georgia Institute of Technology; 2000.

14. Pearson D, editor Global Positioning System (GPS) and travel surveys: Results from the 1997 Austin household survey. Eighth Conference on the Application of Transportation Planning Methods, Corpus Christi, Texas; 2001.

15. Wolf J, Guensler R, Bachman W. Elimination of the travel diary: Experiment to derive trip purpose from global positioning system travel data. Transportation Research Record. 2001;1768(1):125-34.

16. Ojah M, Pearson D. Austin/San Antonio GPS-Enhanced Household Travel Survey,". Texas Transportation Institute. 2008.

17. Wolf J, Lee M, editors. Synthesis of and statistics for recent GPS-enhanced travel surveys. Paper submitted to the Eighth Int Conf Survey Methods in Transport: Harmonization and Data Comparability, Annecy, France; 2008.

18. Wolf J, Oliveira M, Thompson M. Impact of underreporting on mileage and travel time estimates: Results from global positioning system-enhanced household travel survey. Transportation research record. 2003;1854(1):189-98.

19. Council TCM. 2010-2012 Minneapolis - St. Paul Travel Behavior Inventory. 2012.

20. Commission DVRP. 2012-2013 Delaware Valley Household Travel Survey. 2013.

21. Westat R. 2014 Southern Nevada Household Travel. Final Report.; 2015.

22. Shen L, Stopher PR. Review of GPS travel survey and GPS data-processing methods. Transport reviews. 2014;34(3):316-34.

23. Itsubo S, Hato E. Effectiveness of household travel survey using GPS-equipped cell phones and Web diary: Comparative study with paper-based travel survey. 2006.

24. Krygsman SC, Nel J. The use of global positioning devices in travel surveys-a developing country application. SATC 2009. 2009.

25. Stopher P, Wargelin L, editors. Conducting a household travel survey with GPS: reports on a pilot study. 12th World Conference on Transport Research; 2010.

26. Schönfelder S, Axhausen KW, Antille N, Bierlaire M. Exploring the potentials of automatically collected GPS data for travel behaviour analysis: A Swedish data source. Arbeitsberichte Verkehrs-und Raumplanung. 2002;124.

27. Papinski D, Scott DM, Doherty ST. Exploring the route choice decision-making process: A comparison of planned and observed routes obtained using person-based GPS. Transportation research part F: traffic psychology and behaviour. 2009;12(4):347-58.

28. INRIX Traffic, https://inrix.com/ 2021

29. Haghani A, Hamedi M, Sadabadi KF. I-95 Corridor coalition vehicle probe project: Validation of INRIX data. I-95 Corridor Coalition. 2009;9.

30. Schrank D, Eisele B, Lomax T. 2014 Urban mobility report: powered by Inrix Traffic Data. 2015.

31. Horak R. Telecommunications and data communications handbook: John Wiley & Sons; 2007.

32. Pinelli F, Di Lorenzo G, Calabrese F, editors. Comparing urban sensing applications using event and network-driven mobile phone location data. 2015 16th IEEE International Conference on Mobile Data Management; 2015: IEEE.

33. Kang C, Ma X, Tong D, Liu Y. Intra-urban human mobility patterns: An urban morphology perspective. Physica A: Statistical Mechanics and its Applications. 2012;391(4):1702-17.

34. Pappalardo L, Simini F, Rinzivillo S, Pedreschi D, Giannotti F, Barabási A-L. Returners and explorers dichotomy in human mobility. Nature communications. 2015;6(1):1-8.

35. Song C, Qu Z, Blumm N, Barabási A-L. Limits of predictability in human mobility. Science. 2010;327(5968):1018-21.

36. Çolak S, Lima A, González MC. Understanding congested travel in urban areas. Nature communications. 2016;7(1):1-8.

37. Bachir D, Khodabandelou G, Gauthier V, El Yacoubi M, Puchinger J. Inferring dynamic origin-destination flows by transport mode using mobile phone data. Transportation Research Part C: Emerging Technologies. 2019;101:254-75.

38. Fekih M, Bellemans T, Smoreda Z, Bonnel P, Furno A, Galland S. A data-driven approach for origin–destination matrix construction from cellular network signalling data: a case study of Lyon region (France). Transportation. 2021;48(4):1671-702.

39. Williams NE, Thomas T, Dunbar M, Eagle N, Dobra A, editors. Measurement of human mobility using cell phone data: developing big data for demographic science. Population Association of America Annual Meeting; 2013: Citeseer.

40. Frias-Martinez V, Virseda J, Rubio A, Frias-Martinez E, editors. Towards large scale technology impact analyses: Automatic residential localization from mobile phone-call data. Proceedings of the 4th ACM/IEEE international conference on information and communication technologies and development; 2010.

41. Soto V, Frias-Martinez V, Virseda J, Frias-Martinez E, editors. Prediction of socioeconomic levels using cell phone records. International conference on user modeling, adaptation, and personalization; 2011: Springer.

42. Chen C, Ma J, Susilo Y, Liu Y, Wang M. The promises of big data and small data for travel behavior (aka human mobility) analysis. Transportation research part C: emerging technologies. 2016;68:285-99.

43. Wang F, Chen C. On data processing required to derive mobility patterns from passively-generated mobile phone data. Transportation Research Part C: Emerging Technologies. 2018;87:58-74.

44. Wang F, Wang J, Cao J, Chen C, Ban XJ. Extracting trips from multi-sourced data for mobility pattern analysis: An app-based data example. Transportation Research Part C: Emerging Technologies. 2019;105:183-202.

45. Flake L, Lee M, Hathaway K, Greene E. Use of smartphone panels for viable and cost-effective GPS data collection for small and medium planning agencies. Transportation Research Record. 2017;2643(1):160-5.

46. AirSage, https://www.airsage.com/ 2021

47. Huang H, Cheng Y, Weibel R. Transport mode detection based on mobile phone network data: A systematic review. Transportation Research Part C: Emerging Technologies. 2019;101:297-312.

48. Burkhard O, Becker H, Weibel R, Axhausen KW. On the requirements on spatial accuracy and sampling rate for transport mode detection in view of a shift to passive signalling data. Transportation Research Part C: Emerging Technologies. 2020;114:99-117.

49. Gong L, Morikawa T, Yamamoto T, Sato H. Deriving personal trip data from GPS data: A literature review on the existing methodologies. Procedia-Social and Behavioral Sciences. 2014;138:557-65.

50. Axhausen K, Schonfelder S, Wolf J, Oliveria M, Samaga U, editors. Eighty weeks of gps traces, approaches to enriching trip information. Transportation Research Board Annual Meeting; 2004: Citeseer.

51. Stopher PR, Jiang Q, FitzGerald C. Processing GPS data from travel surveys. 2nd international colloqium on the behavioural foundations of integrated land-use and transportation models: frameworks, models and applications, Toronto. 2005.

52. Tsui SYA, Shalaby AS. Enhanced system for link and mode identification for personal travel surveys based on global positioning systems. Transportation Research Record. 2006;1972(1):38-45.

53. McGowen P, McNally M, editors. Evaluating the potential to predict activity types from GPS and GIS data. Transportation Research Board 86th Annual Meeting; 2007: Citeseer.

54. Du J, Aultman-Hall L. Increasing the accuracy of trip rate information from passive multi-day GPS travel datasets: Automatic trip end identification issues. Transportation Research Part A: Policy and Practice. 2007;41(3):220-32.

55. Stopher P, FitzGerald C, Zhang J. Search for a global positioning system device to measure person travel. Transportation Research Part C: Emerging Technologies. 2008;16(3):350-69.

56. Schuessler N, Axhausen KW. Processing raw data from global positioning systems without additional information. Transportation Research Record. 2009;2105(1):28-36.

57. Bohte W, Maat K. Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands. Transportation Research Part C: Emerging Technologies. 2009;17(3):285-97.

58. Gong H, Chen C, Bialostozky E, Lawson CT. A GPS/GIS method for travel mode detection in New York City. Computers, Environment and Urban Systems. 2012;36(2):131-9.

59. Safi H, Assemi B, Mesbah M, Ferreira L. Trip detection with smartphone-assisted collection of travel data. Transportation Research Record. 2016;2594(1):18-26.

60. Patterson Z, Fitzsimmons K. Datamobile: Smartphone travel survey experiment. Transportation Research Record. 2016;2594(1):35-43.

61. Gong L, Sato H, Yamamoto T, Miwa T, Morikawa T. Identification of activity stop locations in GPS trajectories by density-based clustering method combined with support vector machines. Journal of Modern Transportation. 2015;23(3):202-13.

62. Gong L, Yamamoto T, Morikawa T. Identification of activity stop locations in GPS trajectories by DBSCAN-TE method combined with support vector machines. Transportation research procedia. 2018;32:146-54.

63. Zhou C, Jia H, Juan Z, Fu X, Xiao G. A data-driven method for trip ends identification using large-scale smartphone-based GPS tracking data. IEEE Transactions on Intelligent Transportation Systems. 2016;18(8):2096-110.

64. Yao Z, Zhou J, Jin PJ, Yang F. Trip End Identification based on Spatial-Temporal Clustering Algorithm using Smartphone GPS Data. 2019.

65. Yang X, Sun Z, Ban XJ, Holguín-Veras J. Urban freight delivery stop identification with GPS data. Transportation Research Record. 2014;2411(1):55-61.

66. Ye Y, Zheng Y, Chen Y, Feng J, Xie X, editors. Mining individual life pattern based on location history. 2009 tenth international conference on mobile data management: Systems, services and middleware; 2009: IEEE.

67. Calabrese F, Pereira FC, Di Lorenzo G, Liu L, Ratti C, editors. The geography of taste: analyzing cell-phone mobility and social events. International conference on pervasive computing; 2010: Springer.

68. Chen C, Bian L, Ma J. From sightings to activity locations: how well can we guess the locations visited from mobile phone sightings. Transp Res Part C. 2014;46(10):326-37.

69. Zhou C, Frankowski D, Ludford P, Shekhar S, Terveen L. Discovering personally meaningful places: An interactive clustering approach. ACM Transactions on Information Systems (TOIS). 2007;25(3):12-es.

70. Chen W, Ji M, Wang J. T-DBSCAN: A Spatiotemporal Density Clustering for GPS Trajectory Segmentation. International Journal of Online Engineering. 2014;10(6).

71. Yin M. Activity-Based Urban Mobility Modeling from Cellular Data: University of California, Berkeley; 2018.

72. Ester M, Kriegel H-P, Sander J, Xu X, editors. A density-based algorithm for discovering clusters in large spatial databases with noise. kdd; 1996.

73. Jiang S, Fiore GA, Yang Y, Ferreira Jr J, Frazzoli E, González MC, editors. A review of urban computing for mobile phone traces: current methods, challenges and opportunities. Proceedings of the 2nd ACM SIGKDD international workshop on Urban Computing; 2013.

74. Phithakkitnukoon S, Horanont T, Di Lorenzo G, Shibasaki R, Ratti C, editors. Activity-aware map: Identifying human daily activity pattern using mobile phone data. International workshop on human behavior understanding; 2010: Springer.

75. Xie K, Deng K, Zhou X, editors. From trajectories to activities: a spatio-temporal join approach. Proceedings of the 2009 International Workshop on Location Based Social Networks; 2009.

76. Huang L, Li Q, Yue Y, editors. Activity identification from GPS trajectories using spatial temporal POIs' attractiveness. Proceedings of the 2nd ACM SIGSPATIAL International Workshop on location based social networks; 2010.

77. Spinsanti L, Celli F, Renso C, editors. Where you stop is who you are: understanding people's activities by places visited. the proceedings of Behaviour Monitoring and Interpretation (BMI) workshop; 2010.

78. Gong L, Liu X, Wu L, Liu Y. Inferring trip purposes and uncovering travel patterns from taxi trajectory data. Cartography and Geographic Information Science. 2016;43(2):103-14.

79. Flamm M, Kaufmann V. The concept of personal network of usual places as a tool for analysing human activity spaces: a quantitative exploration. Lausanne: EPFL. 2006:23.

80. Calabrese F, Di Lorenzo G, Liu L, Ratti C. Estimating Origin-Destination flows using opportunistically collected mobile phone location data from one million users in Boston Metropolitan Area. 2011.

81. Isaacman S, Becker R, Cáceres R, Kobourov S, Martonosi M, Rowland J, et al., editors. Identifying important places in people's lives from cellular network data. International conference on pervasive computing; 2011: Springer.

82. Yang M, Pan Y, Darzi A, Ghader S, Xiong C, Zhang L. A data-driven travel mode share estimation framework based on mobile device location data. Transportation. 2021:1-45.

83. Stenneth L, Wolfson O, Yu PS, Xu B, editors. Transportation mode detection using mobile phones and GIS information. Proceedings of the 19th ACM SIGSPATIAL international conference on advances in geographic information systems; 2011.

84. Brunauer R, Hufnagl M, Rehrl K, Wagner A, editors. Motion pattern analysis enabling accurate travel mode detection from GPS data only. 16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013); 2013: IEEE.

85. Nitsche P, Widhalm P, Breuss S, Brändle N, Maurer P. Supporting large-scale travel surveys with smartphones–A practical approach. Transportation Research Part C: Emerging Technologies. 2014;43:212-21.

86. Xiao G, Juan Z, Zhang C. Travel mode detection based on GPS track data and Bayesian networks. Computers, Environment and Urban Systems. 2015;54:14-22.

87. Shafique MA, Hato E. Travel mode detection with varying smartphone data collection frequencies. Sensors. 2016;16(5):716.

88. Wang B, Gao L, Juan Z. Travel mode detection using GPS data and socioeconomic attributes based on a random forest classifier. IEEE Transactions on Intelligent Transportation Systems. 2017;19(5):1547-58.

89. Dabiri S, Heaslip K. Inferring transportation modes from GPS trajectories using a convolutional neural network. Transportation research part C: emerging technologies. 2018;86:360-71.

90. Broach J, Dill J, McNeil NW. Travel mode imputation using GPS and accelerometer data from a multi-day travel survey. Journal of Transport Geography. 2019;78:194-204.

91. Vaughan J, Imani AF, Yusuf B, Miller EJ. Modelling cellphone trace travel mode with neural networks using transit smartcard and home interview survey data. European Journal of Transport and Infrastructure Research. 2020;20(4):269-85.

92. Breyer N, Gundlegård D, Rydergren C. Travel mode classification of intercity trips using cellular network data. Transportation Research Procedia. 2021;52:211-8.

93. Group WHOW. Nonpharmaceutical interventions for pandemic influenza, international measures. Emerging infectious diseases. 2006;12(1):81.

94. Brownstein JS, Wolfe CJ, Mandl KD. Empirical evidence for the effect of airline travel on inter-regional influenza spread in the United States. PLoS Med. 2006;3(10):e401.

95. Bajardi P, Poletto C, Ramasco JJ, Tizzoni M, Colizza V, Vespignani A. Human mobility networks, travel restrictions, and the global spread of 2009 H1N1 pandemic. PloS one. 2011;6(1):e16591.

96. Chinazzi M, Davis JT, Ajelli M, Gioannini C, Litvinova M, Merler S, et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. Science. 2020;368(6489):395-400.

97. Kelso JK, Milne GJ, Kelly H. Simulation suggests that rapid activation of social distancing can arrest epidemic development due to a novel strain of influenza. BMC public health. 2009;9(1):1-10.

98. Greenstone M, Nigam V. Does social distancing matter? University of Chicago, Becker Friedman Institute for Economics Working Paper. 2020(2020-26).

99. Li D, Lv J, Botwin G, Braun J, Cao W, Li L, et al. Estimating the scale of COVID-19 epidemic in the United States: Simulations based on air traffic directly from Wuhan, China. MedRxiv. 2020.

100. Koo JR, Cook AR, Park M, Sun Y, Sun H, Lim JT, et al. Interventions to mitigate early spread of SARS-CoV-2 in Singapore: a modelling study. The Lancet Infectious Diseases. 2020;20(6):678-88.

101. Prem K, Liu Y, Russell TW, Kucharski AJ, Eggo RM, Davies N, et al. The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study. The Lancet Public Health. 2020;5(5):e261-e70.

102. Prem K, Cook AR, Jit M. Projecting social contact matrices in 152 countries using contact surveys and demographic data. PLoS computational biology. 2017;13(9):e1005697.

103. Cowling BJ, Ali ST, Ng TW, Tsang TK, Li JC, Fong MW, et al. Impact assessment of non-pharmaceutical interventions against coronavirus disease 2019 and influenza in Hong Kong: an observational study. The Lancet Public Health. 2020;5(5):e279-e88.

104. Bragazzi NL, Dai H, Damiani G, Behzadifar M, Martini M, Wu J. How big data and artificial intelligence can help better manage the COVID-19 pandemic. International journal of environmental research and public health. 2020;17(9):3176.

105. Vaishya R, Javaid M, Khan IH, Haleem A. Artificial Intelligence (AI) applications for COVID-19 pandemic. Diabetes & Metabolic Syndrome: Clinical Research & Reviews. 2020;14(4):337-9.

106. Google. See how your community is moving around differently due to COVID-19. [Available from: https://www.google.com/covid19/mobility/.]

107. Apple. Mobility Trends Reports. 2021 [Available from: https://covid19.apple.com/mobility.]

108. Cuebiq. Mobility Insights. [Available from: https://www.cuebiq.com/visitation-insights-covid19/.]

109. Huang S-K, Lindell MK, Prater CS. Who leaves and who stays? A review and statistical meta-analysis of hurricane evacuation studies. Environment and Behavior. 2016;48(8):991-1029.

110. Murray-Tuite P, Wolshon B. Evacuation transportation modeling: An overview of research, development, and practice. Transportation Research Part C: Emerging Technologies. 2013;27:25-45.

111. Wolshon PB. Transportation's role in emergency evacuation and reentry: Transportation Research Board; 2009.

112. Collier J, Balakrishnan S, Zhang Z. From Hurricane Katrina to Hurricane Harvey: Actions, Issues, and Lessons Learned in Transportation and Logistics Efforts for Emergency Response. 2019.

113. Yin W, Murray-Tuite P, Ukkusuri SV, Gladwin H. An agent-based modeling system for travel demand simulation for hurricane evacuation. Transportation research part C: emerging technologies. 2014;42:44-59.

114. Brown C, White W, van Slyke C, Benson JD. Development of a strategic hurricane evacuation–dynamic traffic assignment model for the Houston, Texas, Region. Transportation research record. 2009;2137(1):46-53.

115. Wang H, Mostafizi A, Cramer LA, Cox D, Park H. An agent-based model of a multimodal near-field tsunami evacuation: Decision-making and life safety. Transportation Research Part C: Emerging Technologies. 2016;64:86-100.

116. Feng K, Lin N. Simulation of hurricane Irma evacuation process. 2019.

117. Mostafizi A, Wang H, Dong S. Understanding the multimodal evacuation behavior for a near-field tsunami. Transportation research record. 2019;2673(11):480-92.

118. Robinson RM, Collins AJ, Jordan CA, Foytik P, Khattak AJ. Modeling the impact of traffic incidents during hurricane evacuations using a large scale microsimulation. International journal of disaster risk reduction. 2018;31:1159-65.

119. Zhang Z, Wolshon B, Herrera N, Parr S. Assessment of post-disaster reentry traffic in megaregions using agent-based simulation. Transportation research part D: transport and environment. 2019;73:307-17.

120. Wong S, Shaheen S, Walker J. Understanding evacuee behavior: a case study of hurricane Irma. 2018.

121. Wu H-C, Lindell MK, Prater CS. Logistics of hurricane evacuation in Hurricanes Katrina and Rita. Transportation research part F: traffic psychology and behaviour. 2012;15(4):445-61.

122. Liu S, Murray-Tuite P, Schweitzer L. Incorporating household gathering and mode decisions in large-scale no-notice evacuation modeling. Computer-Aided Civil and Infrastructure Engineering. 2014;29(2):107-22.

123. Yang H, Morgul EF, Ozbay K, Xie K. Modeling evacuation behavior under hurricane conditions. Transportation research record. 2016;2599(1):63-9.

124. Kontou E, Murray-Tuite P, Wernstedt K. Duration of commute travel changes in the aftermath of Hurricane Sandy using accelerated failure time modeling. Transportation Research Part A: Policy and Practice. 2017;100:170-81.

125. Hasan S, Ukkusuri S, Gladwin H, Murray-Tuite P. Behavioral model to understand household-level hurricane evacuation decision making. Journal of Transportation Engineering. 2011;137(5):341-8.

126. Smith SK, McCarty C. Fleeing the storm (s): An examination of evacuation behavior during Florida's 2004 hurricane season. Demography. 2009;46(1):127-45.

127. Robinson RM, Foytik P, Jordan C. Review and analysis of user inputs to online evacuation modeling tool. 2017.

128. McCarney R, Warner J, Iliffe S, Van Haselen R, Griffin M, Fisher P. The Hawthorne Effect: a randomised, controlled trial. BMC medical research methodology. 2007;7(1):1-8.

129. Groves RM. Survey errors and survey costs: John Wiley & Sons; 2005.

130. Furnham A. Response bias, social desirability and dissimulation. Personality and individual differences. 1986;7(3):385-400.

131. Kumar D, Ukkusuri SV, editors. Utilizing geo-tagged tweets to understand evacuation dynamics during emergencies: A case study of Hurricane Sandy. Companion Proceedings of the The Web Conference 2018; 2018.

132. Roy KC, Hasan S. Modeling the dynamics of hurricane evacuation decisions from twitter data: an input output hidden markov modeling approach. Transportation research part C: emerging technologies. 2021;123:102976.

133. Wang Q, Taylor JE. Quantifying human mobility perturbation and resilience in Hurricane Sandy. PLoS one. 2014;9(11):e112608.

134. Yabe T, Tsubouchi K, Fujiwara N, Sekimoto Y, Ukkusuri SV. Understanding post-disaster population recovery patterns. Journal of the Royal Society Interface. 2020;17(163):20190532.

135. Batini C, Cappiello C, Francalanci C, Maurino A. Methodologies for data quality assessment and improvement. ACM computing surveys (CSUR). 2009;41(3):1-52.

136. Zhang L, Darzi A, Ghader S, Pack ML, Xiong C, Yang M, et al. Interactive covid-19 mobility impact and social distancing analysis platform. Transportation Research Record. 2020:03611981211043813.

137. Bureau. USC. 2019 American Community Survey (ACS) 5-Year Estimates Table DP05: ACS Demographic and Housing Estimates. [Available from: https://data.census.gov/cedsci/all?q=acs.]

138. Bureau USC. Longitudinal-Employer Household Dynamics (LEHD) Origin-Destination Employment Statistics Data (2019) [Available from: https://lehd.ces.census.gov/data/#lodes.]

139. SafeGraph. Home Algo v1 "Monthly Batched" [Available from: https://docs.safegraph.com/docs/monthly-patterns#section-algorithms.]

140. Bureau. USC. 2019 American Community Survey (ACS) 5-Year Estimates Table DP03: Selected Economic Characteristics. August 2021. [Available from: https://data.census.gov/cedsci/all?q=acs.]

141. Bureau. USC. 2011-2015 5-Year American Community Survey (ACS) Commuting Flows, Table 1. Residence County to Workplace County Commuting Flows for the United States and Puerto Rico Sorted by Residence Geography. [Available from: https://www.census.gov/data/tables/2015/demo/metro-micro/commuting-flows-2015.html.]

142. Graham MR, Kutzbach MJ, McKenzie B. Design comparison of LODES and ACS commuting data products. 2014.

143. Xiong C, Darzi A, Pan Y, Ghader S, Zhang L. A Data-Driven Analytical Framework of Estimating Multimodal Travel Demand Patterns using Mobile Device Location Data. arXiv preprint arXiv:201204776. 2020.

144. Géron A. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems: O'Reilly Media; 2019.

145. Duchi J, Hazan E, Singer Y. Adaptive subgradient methods for online learning and stochastic optimization. Journal of machine learning research. 2011;12(7).

146. Tieleman T, Hinton G. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural networks for machine learning. 2012;4(2):26-31.

147. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980. 2014.

148. Pan Y, Darzi A, Kabiri A, Zhao G, Luo W, Xiong C, et al. Quantifying human mobility behaviour changes during the COVID-19 outbreak in the United States. Scientific Reports. 2020;10(1):1-9.

149. Qualls N, Levitt A, Kanade N, Wright-Jegede N, Dopson S, Biggerstaff M, et al. Community mitigation guidelines to prevent pandemic influenza—United States, 2017. MMWR Recommendations and Reports. 2017;66(1):1.

150. Wood HO, Neumann F. Modified Mercalli intensity scale of 1931. Bulletin of the Seismological Society of America. 1931;21(4):277-83.

151. U.S. News & World Report. Best states 2021: how they were ranked. [Available from: https://www.usnews.com/news/best-states/articles/methodology.]

152. World Population Review. Healthiest Countries Population 2021. [Available from: https://worldpopulationreview.com/country-rankings/healthiest-countries.]

153. Federal Highway Administration. 2017 National Household Travel Survey Travel Profile: United States. (U.S. Department of Transportation, Washington, D.C.) [Available from: https://nhts.ornl.gov/assets/2017_USTravelProfile.pdf.]

154. Johns Hopkins University. COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. [Available from: https://coronavirus.jhu.edu/map.html.]

155. Darzi A, Frias-Martinez V, Ghader S, Younes H, Zhang L. Constructing Evacuation Evolution Patterns and Decisions Using Mobile Device Location Data: A Case Study of Hurricane Irma. arXiv preprint arXiv:210212600. 2021.

156. Younes H, Darzi A, Zhang L. How effective are evacuation orders? An analysis of decision making among vulnerable populations in Florida during hurricane Irma. Travel behaviour and society. 2021;25:144-52.

157. Csáji BC, Browet A, Traag VA, Delvenne J-C, Huens E, Van Dooren P, et al. Exploring the mobility of mobile phone users. Physica A: statistical mechanics and its applications. 2013;392(6):1459-73.

158. POLL M-DF. Hurricane Irma 2017 [Available from: https://media.news4jax.com/document_dev/2017/10/26/Mason-Dixon%20Hurricane%20poll_1509043928726_10861977_ver1.0.pdf.]

159. Webster PJ, Holland GJ, Curry JA, Chang H-R. Changes in tropical cyclone number, duration, and intensity in a warming environment. Science. 2005;309(5742):1844-6.

160. Elsner JB, Kossin JP, Jagger TH. The increasing intensity of the strongest tropical cyclones. Nature. 2008;455(7209):92-5.