

ABSTRACT

Title: ENHANCING DECISION-MAKING IN SMART
AND CONNECTED COMMUNITIES
WITH DIGITAL TRACES

Lingzi Hong
Doctor of Philosophy, 2019

Dissertation directed by: Assistant Professor Vanessa Frias-Martinez
College of Information Studies

The ubiquitous use of information communication technologies (ICTs) enables generation of digital traces associated with human behaviors at unprecedented breadth, depth, and scale. Large-scale digital traces provide the potential to understand population behaviors automatically, including the characterization of how individuals interact with the physical environment. As a result, the use of digital traces generated by humans might mitigate some of the challenges associated to the use of surveys to understand human behaviors such as, high cost in collecting information, lack of quality real-time information, and hard to capture behavioral level information. In this dissertation, I study how to extract information from digital traces to characterize human behavior in the built environment; and how to use such information to enhance decision-making processes in the area of Smart and Connected Communities. Specifically, I present three case studies that aim at using data-driven methods for decision-making in Smart and Connected Communities. First, I discuss data-driven methods for socioeconomic development with a

focus on inference of socioeconomic maps with cell phone data. Second, I present data-driven methods for emergency preparedness and response, with a focus on understanding user needs in different communities with geotagged social media data. Third, I describe data-driven methods for migration studies, focusing on characterizing the post-migration behaviors of internal migrants with cell phone data. In these case studies, I present data-driven frameworks that integrate innovative behavior modeling approaches to help solve decision-making questions using digital traces. The explored methods enhance our understanding of how to model and explain population behavior patterns in different physical and socioeconomic contexts. The methods also have practical significance in terms of how decision-making can become cost-effective and efficient with the help of data-driven methods.

ENHANCING DECISION-MAKING IN SMART AND
CONNECTED COMMUNITIES WITH DIGITAL TRACES

by

Lingzi Hong

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
2019

Advisory Committee:

Assistant Professor Vanessa Frias-Martinez, Chair/Advisor

Professor Richard Marciano

Associate Professor Hiroyuki Iseki

Assistant Professor Grant McKenzie

Assistant Professor Hernisa Kacorri

© Copyright by
Lingzi Hong
2019

To my parents

Acknowledgments

I am really thankful to all the people who have made this thesis possible. Without their support and encouragement, I can't imagine how to achieve this. Thanks to them for they have made my graduate life worth cherishing.

First, I would like to thank my advisor, Vanessa Frias-Martinez. She gave me the opportunity to work with her on really interesting and challenging projects over the five years. She provided a lot of guidance throughout my graduate life, not only in intellectual perspective but also in the altitude for life. From her I got to know although achievement in research is critical, it should never be the priority or the source of pressure. It is important to enjoy the working process and make research work an exciting and inseparable part of our life.

I would also like to thank my committee: Richard Marciano, Hiroyuki Iseki, Grant McKenzie, and Hernisa Kacorri for their time reviewing the manuscript. Thanks for their challenges, guidance and support that help me complete this work.

I received a lot of help from my colleagues. Special thanks to Jiahui Wu, whom I have worked with in many projects such as the migration study and cycling safety project. I really enjoyed the time when we discussed research ideas and had intensive arguments. It helped train my research mindset and find ways out for tough questions. Thanks to Myeong Lee, Cheng Fu, Jiaying He, Weiwei Yang and many others who I have collaborated with. They have helped me develop this work into what it is now. Thanks to my cohorts Leyla Norooz, Brenna McNally who gave me encouragement when I felt difficulties and had doubts in pursuing my degree.

My family and my friends deserve a special mention for they have been very supportive and encouraging. I always feel love from them, which is the force that enable me to embrace challenges and keep pursuing my goal. Thanks to my parents and my sister. We were not be able to stay together most of the time but we were always connected. Words cannot express the gratitude I owe them. Thanks to Yurong He, Yuting Liao, Yacong Yuan, Yuyu Wang, Zhenpeng Zhao, Xuyang Zhou, Jun Li and many friends who shared part of their time with me. It is a too long list to be fully enumerated. I will keep the invaluable memory in my mind.

And of course a very special thanks to my husband, Dong Nie. He was a PhD student in Computer Science at the University of North Carolina Chapel Hill. He always stood by me and supported me through my career. His love is the reason for me to be strong and go through all the difficulties. He is also the role model of my career.

It is impossible to remember all, and I apologize to those I've inadvertently left out.

Lastly, thank you all and thank God! For God did not give us a spirit of timidity, but a spirit of power, of love and of self-discipline (2 Timothy 1:7).

Table of Contents

| | |
|--|------|
| Dedication | ii |
| Acknowledgements | iii |
| List of Tables | viii |
| List of Figures | ix |
| 1 Introduction | 1 |
| 1.1 Contribution | 5 |
| 1.2 Organization | 7 |
| 2 Motivation and Research Questions | 10 |
| 2.1 Challenges of Decision-making in Smart and Connected Communities | 11 |
| 2.2 Large-scale Digital Traces | 13 |
| 2.2.1 Generation of Digital Traces | 14 |
| 2.2.2 Making Sense of Population Behaviors Using Digital Traces | 15 |
| 2.2.3 Characteristics of Large-scale Digital Traces | 18 |
| 2.3 Data-driven Methods for Three Research Areas | 20 |
| 2.3.1 Socioeconomic Development | 20 |
| 2.3.2 Emergency Preparedness and Response | 21 |
| 2.3.3 Migration Studies | 22 |
| 2.4 Challenges of Developing Data-driven Methods | 24 |
| 2.4.1 Identification of Research Subjects | 24 |
| 2.4.2 Identification of Behavioral Features | 26 |
| 2.4.3 Identification of Modeling Approaches | 26 |
| 3 Literature Review | 29 |
| 3.1 Data-driven Methods for Socioeconomic Development | 29 |
| 3.1.1 Analytic Strategies for Inference of Socioeconomic Development | 31 |
| 3.1.2 Behavioral Features Relating to Socioeconomic Development | 32 |
| 3.2 Data-driven Methods for Emergency Preparedness and Responses | 34 |
| 3.2.1 Digital Traces Used in Different Phases of Disasters | 34 |
| 3.2.2 Understanding Online Communications for Disaster Response | 36 |

| | | |
|-------|---|-----|
| 3.3 | Data-driven Methods for Migration Study | 40 |
| 3.3.1 | Migration Models Built on Crowdsourced Data | 41 |
| 3.3.2 | Survey-based Migration Analyses | 42 |
| 4 | Study 1: Inferring Socioeconomic Maps with Topic Models on Mobility Motifs | 44 |
| 4.1 | Research Question | 45 |
| 4.2 | The Proposed Method | 46 |
| 4.2.1 | PF | 47 |
| 4.2.2 | PF2 | 50 |
| 4.2.3 | PMB-LDA | 51 |
| 4.2.4 | PMB-sLDA | 53 |
| 4.3 | Results | 55 |
| 4.3.1 | Dataset | 55 |
| 4.3.2 | SEL Inference | 56 |
| 4.4 | Discussions | 59 |
| 4.5 | Limitations | 60 |
| 5 | Study 2: Understanding Online Communications during Natural Disasters | 62 |
| 5.1 | Research Question | 62 |
| 5.2 | The Proposed Method | 64 |
| 5.2.1 | Data Preparation | 66 |
| 5.2.2 | Tweet Ranking and Selection | 68 |
| 5.2.3 | Topic Identification | 71 |
| 5.3 | Results | 74 |
| 5.3.1 | Spatiotemporal Burst | 74 |
| 5.3.2 | Themes and Topics in Roads | 78 |
| 5.3.3 | Interactions between Citizens and Local Government Accounts | 80 |
| 5.4 | Limitations | 83 |
| 6 | Study 3: Identification and Characterization of Internal Migrants with Cell Phone Data | 84 |
| 6.1 | Introduction | 84 |
| 6.2 | The Proposed Method | 86 |
| 6.3 | Data Description | 89 |
| 6.3.1 | Cell Phone Data | 89 |
| 6.3.2 | Census Data | 90 |
| 6.4 | Identification and Validation of Internal Migrants | 91 |
| 6.4.1 | Identification of Internal Migrants | 91 |
| 6.4.2 | Validation of Internal Migrants | 93 |
| 6.5 | Results | 97 |
| 6.5.1 | RQ1: Behavior Consequences of Internal Migrants | 97 |
| 6.5.2 | RQ2: Role of Pre-migration Behaviors on Post-migration Ac- tivities | 109 |
| 6.6 | Potential Limitations | 114 |

| | | |
|-------|--|-----|
| 7 | Conclusions and Future Directions | 117 |
| 7.1 | Conclusions | 117 |
| 7.1.1 | Topic Models to Infer Socioeconomic Maps | 117 |
| 7.1.2 | Understanding Online Communications of Citizens and Local Governments during Natural Disasters | 119 |
| 7.1.3 | Identification and Characterization of Internal Migrants with Cell Phone Data | 121 |
| 7.2 | Limitations | 123 |
| 7.3 | Future Directions | 125 |
| 7.3.1 | Validation in Other Scenarios | 126 |
| 7.3.2 | Development of Modeling Methods | 127 |
| | Bibliography | 130 |

List of Tables

| | | |
|-----|--|-----|
| 4.1 | Accuracies for regression with topic models and pre-determined features. | 58 |
| 4.2 | Accuracy(ACC), average F1 and per-class F1 score with topic models and pre-determined features. | 59 |
| 5.1 | Themes and topics from citizen tweets, * indicates common topics for both citizens and local governments. | 72 |
| 5.2 | Themes and topics from local governments tweets, * indicates common topics for both citizens and local governments. | 73 |
| 5.3 | The @ behavior of citizens under various topics | 81 |
| 6.1 | Statistical analysis of spatial dynamics' features: comparison between migrant and local behaviors. | 99 |
| 6.2 | Statistical analysis of social ties' features: comparison between migrant and local behaviors using Welch's t-test with Cohen's d and Box-Cox transformation for skewed distributions. | 107 |
| 6.3 | Multivariate regressions on pre-migration behavioral features to quantify their role on post-migration behaviors. Adjusted R-squared values are used to assess the predictability of post-migration behaviors with pre-migration features. | 112 |

List of Figures

| | | |
|-----|---|----|
| 1.1 | Structure of this work. | 9 |
| 4.1 | Approach overview: ①PF ②PF2 ③PMB-LDA ④PMB-sLDA. LDA and sLDA plate notation from [1] | 47 |
| 4.2 | R^2 per number of topics for PMB-LDA (SVR and RF) and PMBSEL-sLDA approaches. | 57 |
| 5.1 | Overview of the proposed framework. | 65 |
| 5.2 | PMI values for various thresholds (th) and number of topics (K) and for the keywords approach. | 70 |
| 5.3 | Spatiotemporal clusters detected by STPS. | 75 |
| 5.4 | Themes of tweet clusters by citizens and tweets by local governments in the same spatiotemporal scale. | 76 |
| 5.5 | Distribution of themes and topics on roads. | 80 |
| 6.1 | Overview of the proposed framework and the two research questions it answers: Identification of migrants, (RQ1) Analysis of the behavioral consequences of internal migrants and (RQ2) Analysis of the potential causes of the post-migration behaviors. RQ1 and RQ2 use two types of behavioral traits extracted from the data: spatial dynamics and social ties | 88 |
| 6.2 | Goodness of fit with census-based and CDR-based Migration Flows. The migration models explored are GravExp, NGravExp, GravPow, NGravPow, Schneider, Rad and RadExt. The constraints are UM, PCM, ACM and DCM. | 94 |

Chapter 1: Introduction

The expanded use of digital devices brings an explosive growth of digital traces. Digital devices such as cell phones are widely used even in developing countries. Recent reports show that there are almost 7 billion mobile phone subscriptions at the global level, with three quarters from developing countries [2]. The high penetration of digital devices is associated to the generation of personal data at large scales. Digital devices are used in all aspects of human life, including but not limited to communication with other people, information sharing and services, and navigation and interaction in the physical or built-in environment (e.g. location services). When users interact with digital devices, data about the interaction details are automatically created and archived. Digital devices have become an inseparable part of work, personal life, and public services for more and more people. This leads data to be generated with unprecedented scales, breadth and depth.

Digital traces record different types of information with respect to human behaviors. Digital traces are generated when people communicate with other people, therefore the communication details are recorded, from which we can extract the social relations. People also share information or express opinions online, which generates data that reflect their needs, sentiments or attitudes [3]. Digital devices

such as smartphones with GPS technology, collect information on spatial locations as time series, which reflects mobility activities in the physical environment.

Large scale digital traces have the potential to reveal how people interact with the physical or built-in environment, which is valuable information for certain decision-making processes. Understanding population behaviors in space and time is critical in many decision-making processes, such as the planning and management of facilities [4], identifying targets for aid efforts [5], or disaster response and management [6]. For example, in traffic management, decision-makers have the need to understand population mobility patterns in normal days and when there are events so that they can schedule plans accordingly for traffic flow control [7]. Many researchers and practitioners also hold the opinion that services based on the evidence of human reactions are more efficient [6].

One might argue that traditional survey methods can be used to provide information about human interactions with the built environment that might be of interest for decision-making processes. However, there are limitations if one solely relies on survey data. First, surveys usually require a lot of human effort to distribute questions, collect and process data. Since surveys come at a high cost, usually a limited number of samples are identified to represent a large population, which might lead to coarse information that has poor spatial and temporal granularity [8]. Second, it is hard to collect timely information using survey methods since they are time-consuming. In situations like emergency responses, timely information is helpful to identify the priority of user needs.

Digital traces, on the other hand, can complement the limitations of survey

data. Unlike survey data, digital traces are usually automatically generated when users interact with the digital devices. With a large-scale user group, digital traces such as cell phone records or tweets are continuously generated from a wide spatial scope. Digital traces are more cost-efficient compared to survey data. Digital traces provide information that is hard to obtain by survey data, i.e. detailed behavioral information and real-time information, which can be helpful in situations where such information is needed. Digital traces are a passive way to collect data as opposed to surveys which require active response from survey takers, which means that digital traces present less subjective bias from survey takers.

Previous studies have explored the potential use of digital traces for decision making. Digital traces have been used for situational awareness during natural disasters, which leads to efficient disaster response [9]. Digital traces that are generated by online communications have been proved to support unprecedented levels of mutual government-citizen understanding, and in turn, largely improve public policies and services [10]. Communication data makes the governments engage with their citizens more effectively and actively, and improve their services in health, transportation, energy and many perspectives [11].

Decision makers have become increasingly interested in leveraging the power of data for decision-making processes to improve the overall quality of life. This dissertation is centered around decision making for Smart and Connected Communities, where communities refer to geographically-delineated units that consist of people, built or natural environment and people engaging with the environment. The concept of Smart and Connected Communities refers to communities that inte-

grate intelligent systems that can improve ‘the social, economic and environmental well-being of those who live, work, or travel within it’ [12].

This dissertation works towards innovative approaches of using large-scale digital traces to enhance decision-making processes in Smart and Connected Communities. In this work, digital traces are specifically referred to data that are collected from individual users and have spatiotemporal information embedded. Many data sources can be seen as digital traces. For example, the call detail records (CDRs), which are cell phone metadata that contains spatiotemporal information when the user makes a connection to cellular towers. Geotagged social media data are also digital traces. Digital traces contain signals for us to understand how people react to the changes in the environment from the perspectives of social activities, mobility, and opinion expression.

The vision is that by modeling human behaviors with large-scale digital traces, we can sense and characterize users’ behaviors in terms of spatial, social activities and communication behaviors, to evaluate the effect of socioeconomic context, social events or natural shocks on communication and mobility behaviors. This information helps to enhance the understanding of peoples’ reactions and why they react in a certain way, thus potentially assisting the decision making processes of governments and organizations. Compared to surveys and studies, data-driven methods are usually automatic and scalable, which enables more efficient decision-making.

1.1 Contribution

In the following, I present a summary of contributions of using data-driven methods to enhance decision-making for Smart and Connected Communities in three areas, which are data-driven methods for socioeconomic development, data-driven methods for emergency preparedness and response, and data-driven methods for migration studies. In these projects, I build statistical and machine learning models to extract critical information regarding the mobility, social connections and communications of people, to form the knowledge of behavior prediction, context inference, and impact evaluation. This work contributes with innovative ideas on human behavior modeling and on data-driven frameworks for decision-making in Smart and Connected Communities. Through this work, I show that large-scale digital traces can be used to support decision-making processes, therefore can potentially help to improve public policies and services in an efficient manner.

- **Data-driven methods for Socioeconomic Development:** I present a novel approach to model regional socioeconomic maps with population mobility behaviors that are extracted from cell phone records. Specifically, I define the individual transition activity between regions as mobility motif and use topics models to learn the latent recurring patterns of co-occurring behaviors across regions for the prediction of socioeconomic levels. The approach improves the state of the art prediction results by about 9% and helps to reveal mobility patterns that are indicative of different socioeconomic levels.

Socioeconomic maps are critical to support informed policy-making. However, traditional survey data are usually costly and lead to coarse or delaying maps that affect efficient decision-making. Compilation of socioeconomic maps with digital traces can complement survey information in a cost-effective manner, which leads to maps with higher spatial and temporal resolutions. Such kind of maps enable more targeted and effective aid efforts.

- **Data-driven methods for Emergency Preparedness and Response.**

I use social media data (i.e. tweets) to model the needs of citizens during natural disasters and the communication behaviors between citizens and local governments. I propose a semi-automatic framework that requires less human efforts in understanding massive tweet data. The framework is composed of two major parts: a) a language model to automatically evaluate the relevance of tweets to disasters, and b) topic models to learn the digital communication footprints (topics) of citizens and local governments. With these identified digital footprints, I analyze the topics at various spatio-temporal scales and the interactive communication patterns between citizens and local government accounts.

The framework can help to identify the issues citizens are most concerned in online communications to enhance situation awareness. Compared to previous methods, it requires minimal efforts in labeling. The data-driven framework does not incorporate prior information about disasters. Thus, it is applicable for other emergency situations.

- **Data-driven methods for Migration Studies.** I propose a data-driven framework that identifies internal migrants and characterizes their post-migration behaviors. I implement methods to identify migrants based on their spatiotemporal trajectories. The detected migration flow is highly correlated with census migration data. The proposed framework allows to carry out micro-level analyses of internal migrants from two perspectives: spatial behaviors and social ties.

The data-driven results may guide or complement the research agenda of micro-level migration, and enhance the understanding of the physical, social and psychological decision processes behind migration experiences.

1.2 Organization

The structure of the dissertation is as follows:

In Chapter 1, I first introduce the background. We enter a new era where digital devices facilitate communications between people and people with the physical environment. Large scale digital traces of behavioral information with fine granularity are generated. I argue that digital traces have the potential to enhance our understanding of the interaction between people and the environment, which is useful information for the improvement of public policies and services. In this work, I mainly use digital traces with spatiotemporal information. The dissertation works towards the modeling of digital traces to assist decision-making in Smart and Connected Communities. Then I summarize the contribution of this work to three

areas, including the intellectual and societal impacts.

Chapter 2 presents the motivations for this work. There are many challenges of decision-making in communities, which lead to inefficiency in public services. Traditional survey methods usually require extensive resources and have time lag in providing evidence for decision-making. Digital traces become a valuable source of information to understand human behaviors. From digital traces, we can model the mobility behaviors, social networks, and online communications at the individual and population level. Digital traces provide information that is hard to achieve through traditional survey methods.

There are challenges and limitations of applying data-driven methods in decision-making since the digital traces are generated with complicated mechanism and affected by many factors. It usually requires a re-purposing process to make sense of the data. I look into data-driven solutions for research questions in three areas. In the following chapters I address these challenges and present the new methods.

Chapter 3 summarizes previous studies related to the three case studies. Chapter 4, 5 and 6 present the details of the methods and the experiment results.

In the last chapter, I conclude with the case studies and discuss the ethical issues in data-driven methods, mostly focused on bias and privacy issues. Then I show the future directions to continue and extend this work.

Fig. 1.1 shows the framework of this work. Chapter 1 and Chapter 2 describe the background, generation of digital traces and the characteristics of digital traces. The blue modules show the logic of data-driven processes to enhance decision-making. In Chapter 4,5,6, I present data-driven methods that take into

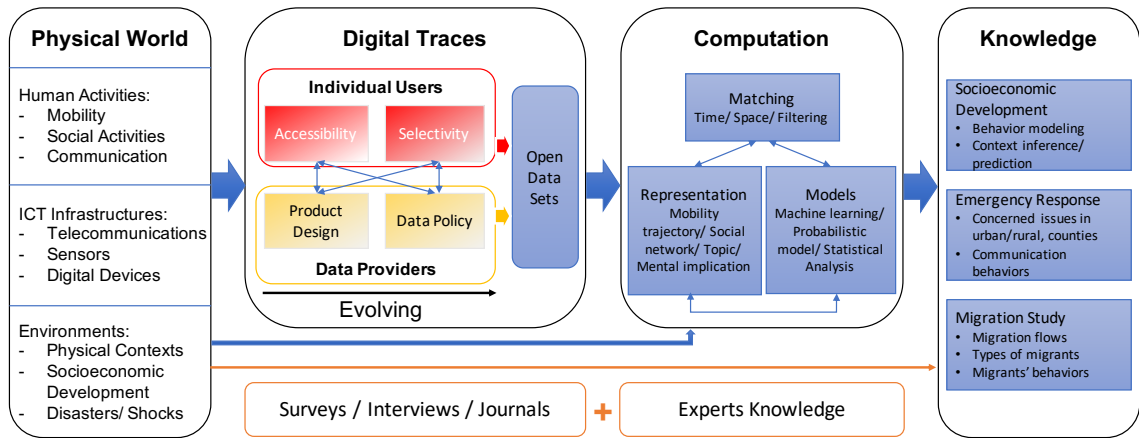


Figure 1.1: Structure of this work.

account the challenges, which are the identification of subjects, behavior features and modeling approaches for three research questions. Data-driven models have the potentials to reveal patterns and relations of how people interact with the surrounding environment, which can complement the current survey methods and lead to enhanced understanding of research questions in the areas of socioeconomic development, emergency preparedness and response and migration studies.

Chapter 2: Motivation and Research Questions

In this dissertation, I am interested in exploring how large-scale digital traces can be used to enhance decision-making processes in Smart and Connected Communities and overcome some of the challenges faced when using traditional survey methods; and in the design of data-driven methods to extract insights from those digital traces. Since the area of Smart and Connected Communities is quite broad, I will focus my work on three areas: socioeconomic development, emergency preparedness and response, and migration studies. In section 2.1, I describe the challenges of decision-making in Smart and Connected Communities with survey methods. Section 2.2 introduces how the large-scale digital traces are generated, what kind of behavior information is embedded in digital traces and the characteristics of digital traces. Section 2.3 presents three specific areas of Smart and Connected Communities I focus on. In section 2.4, I show the challenges of developing data-driven methods to enhance decision-making.

2.1 Challenges of Decision-making in Smart and Connected Communities

Decision-making in communities covers a wide range of issues, such as policy-making, planning, management of facilities or disaster response. Traditionally, decision-making has been implemented using a top-down paradigm whereby policies are put in place and then evaluated [10]. However, there exists an increasing shift towards evidence-based decision making where governments and cities first attempt to understand actual citizens' needs, preferences and behaviors via digital traces, and then propose specific policies to assist those needs [13]. This approach is mostly limited by the data that can be collected via surveys to inform policies in Smart and Connected Communities. In this section, I describe the limitations that traditional survey approaches have, and the next sections will cover digital traces, the information that can be extracted from those, and how these might be used to enhance decision making processes.

High cost and limited samples. Survey methods and interviews have been widely used to collect information for decision-making. Although nowadays surveys and interviews can be largely mediated through computers, they still require human effort to distribute surveys and process survey results [14]. Additionally, many communities have limited budgets that prevent them from sampling with a high temporal frequency or for a large number of individuals, which might lead to biased information for decision-making. An example is the compilation of socioeconomic

maps that depict the economic development status of different regions. Government and non-governmental Organizations (NGOs) rely on the socioeconomic development status to allocate resources, however, the development status keeps changing. Given a limited budget, maps are only computed every several years, which leads to socioeconomic information with poor spatial and temporal granularity [15].

Lack of timely information. Survey methods fail in situations when timely information is required for decision making, for example in the scenario of disaster response. Quick and efficient response is required in these situations, however, it is hard to acquire real-time information to evaluate current status. Although residents may directly contact government agencies that are in charge, there may not be enough people to receive and process these messages, which might cause government responses to be delayed. Piecemeal information is not sufficient for decision-makers to have a macro understanding of the current situation, therefore affecting the efficiency of resource allocation [16]. It requires an understanding of residents' needs and concerns at a large-scale to decide where the communities are most affected, and what are the priority issues to solve. Answers to these questions can differ by disasters and by cases. Decision-making relying on previous experiences therefore can be improved with real-time information for situation awareness. An accurate and in-time understanding of states after disasters is critical for the efficient response to reduce damage and help people in need [17].

Behavioral information is hard to capture. Behavioral information can enhance decision-making. As Conte said, in the background that people are globally connected, the effect of individual behaviors is expanding and might lead to unex-

pected results. It is urgent to have a comprehensive understanding of how people react and connect in the global society [18]. However, it is hard to obtain behavioral information through surveys, since behaviors are a continuous and dynamic process, which involve many details. Take mobility information as an example. Survey data cannot fully capture the real trajectories. Survey takers can describe the trip purposes, estimate the miles traveled, drawing mental maps to recover their mobility activities between friends and families, however, it is almost impossible to achieve a detailed description of trips [19]. Alternatively, digital traces with spatiotemporal information record the exact locations when a real event happens, therefore can retrieve more detailed information on trips.

It is also hard to capture globally interconnected information using sampling methods. For example, survey methods can be used to study ego networks, which are network relations around one person, but not used to study global networks, where everyone is represented in the network, due to the large scale.

2.2 Large-scale Digital Traces

In the digital era, ICTs help to build connections between people, communities and the physical environment. Digital traces are generated and recorded, providing opportunities to model online and offline connections of people and to make sense of population behaviors. Ultimately, such behaviors can be used to inform and enhance decision making processes in Smart and Connected Communities.

2.2.1 Generation of Digital Traces

The development of ICTs has changed the way people communicate, which now heavily depends on digital devices and applications [20]. Digital devices and applications like cell phones and social media have become an inseparable part of how people work, communicate and live. For example, cell phones are heavily used in private settings [21]. People make phone calls to family, friends, work or service related people [22]. Social media sites are also actively used for sharing information and building social connections [23]. In return, one's social relations are embedded in the digital traces left behind.

Digital traces are generated and recorded when people interact with digital devices. Some are automatically and passively generated. For example, CDRs are generated when a user makes a phone call or has connections with nearby cellular towers. They are originally collected by telecommunication companies for billing purposes. Others are initiated by users' check-in or post behaviors. Some applications enable location services, with which users help to create the volunteered geographic information (VGI). Unlike cell phone data, which is a by-product of users' behaviors, these digital traces are actively created by users. Active users may represent only a small proportion of the population. However, the group of people that can be observed might be large and geographically wide.

Cell phone data are held by the telecommunication companies. Some of the data are shared with researchers or through open data projects [24]. Social media data are kept by social media companies, which can be bought or streamed by

APIs [25].

2.2.2 Making Sense of Population Behaviors Using Digital Traces

Digital traces not only record users' interactions in cyberspace but also reflect their behaviors in the physical world. Digital traces provide quality observations with high level of detail and refinement for behavior modeling, which can in turn inform decision making processes in Smart and Connected Communities [22]. Next, I describe the three types of behavior I work with in my dissertation.

Mobility. Mobility activities can be inferred from data that contains geographic information in a time series format. The accuracy of the inference usually depends on the spatial and temporal granularity of the data. Here I mainly discuss the inference of mobility behaviors from CDRs and geotagged social media data.

People carry cell phones in their daily movement. Whenever they make phone calls, use wireless Internet or have other types of connection to the telecommunication network, the nearby cellular towers transfer signals and generate records. Since towers can be identified by the latitude and longitude of their geographic locations, we can infer the approximated locations of cell phone users. For simplicity, it is common to assume that the cell of each tower is a 2-dimensional non-overlapping polygon, which is approximated using Voronoi diagrams. Therefore, the spatial granularity for identification of users' locations depends on the size of cells. It is usually less than $1km^2$ in urban dense areas and more than $4km^2$ in rural areas [26]. Spatial information in time series records can be used to infer mobility trajectories.

The accuracy of trajectory inference is determined by the frequency of cell phone connections to cellular towers.

Applications such as Twitter, Foursquare or Flickr provide location services, which also enable users to share their locations at different geographic scales, i.e. GPS points or names of places. Tweets with latitude and longitude provide accurate location information of users. Frequent geotagged tweets, therefore, can be used to infer users' mobility trajectory. Roughly 1% of all tweets contain latitude and longitude information [27].

Places are also helpful to identify county-level, state-level or country-level locations of users, therefore, can be used to approximate mobility across different regions [28]. Since users are not required to use one standard location identifier system, they tend to label one place with variations of names, which makes it difficult to identify the exact location. However, geographic information at larger geographical areas (*e.g.*, city or country) is more consistently applied among users.

Considering privacy issues, the studies of mobility behaviors are usually at the population level. Previous studies have used anonymized mobile phone records to understand mobility patterns in urban space [29], explore the home-work commuting patterns [30], and identify the laws that govern the regularity in human trajectories [7]. Geotagged tweets have been used to capture cross-border movement [31], and estimate the international and internal migration patterns [32].

Social Networks. ICTs facilitate communications between people and generate digital traces that reflect online and offline social relations. People mostly use cell phones to connect with people whom they have physical contact with, for

example, relationships of work, family, leisure and services [33]. Although cell phone activities also include the interactions with spam numbers or business service numbers, these contacts can be differentiated based on reciprocal activities [34]. In social media sites, digital traces reflect a mix of online and offline social relations. Twitter is perceived and used differently by users. Some use it as a social networking site to connect with friends or families, while some use it as a public platform for brand marketing or a media outlet to release news [23]. Different intentions lead to different behaviors, which regarding whom they choose to interact with. Users also participate in online communities by using Twitter Lists to exchange information with people they don't know yet. The online communities are bridges for people to expand their social networks [35].

Social networks that are built by phone-based activities, or social media activities enable us to infer or estimate one's offline social networks, which in Sociology represents social capital [36]. Social capital is associated with one's socioeconomic level, available resources, social integration, and many other sociological concepts. Online social relations are also critical for understanding people's reactions to disasters or social events. It has been studied that online social relationships are related to the participation in offline events, such as activist movements [37], community support for disaster response [38] and security support [39].

Communications. Social media sites provide online spaces for people to share information and express opinions. Some record personal life or share news on such platforms, using social media sites for networking. While others use them as media outlets to amplify the effect of information dissemination [23]. The com-

munication contents they post online reflect their concerns, stances, and attitudes. These digital traces act as voluntarily contributed responses by users, although it usually requires efforts to re-purpose according to certain research questions.

Researchers have implemented methods to understand the content of communications from different perspectives. Topic models have been widely used to explore topics in social media communication, for example, to identify the trend of topics online [40], to compare topics on social media sites and traditional media [41], and model user interest [42]. Topics are clusters of words that are organized by semantic similarities. Researchers have also looked into the sentiment of communications [43], or mood and emotions events [44]. Dictionaries such as the Linguistic Inquiry and Word Count (LIWC) have been used to exploit the psychological meaning of communications [45]. Words and sentence structures are used to identify the key elements, thus the agents, actions, and targets in the stories told by social media users [46].

2.2.3 Characteristics of Large-scale Digital Traces

The unique characteristics of digital traces enable us to use them to complement the limitations of survey data, therefore enhancing decision-making processes in Smart and Connected Communities. Large-scale digital traces are:

Revealed behaviors. Digital traces are generated by users when they interact with digital devices. Unlike survey methods, where the collection of information requires survey takers to provide "stated" information, digital traces are "revealed" behaviors, automatically collected by service providers, such as telecommunication

companies or social media companies. Digital traces are objective records that reveal users' interactions with the physical world, which might enhance the information provided through surveys that ask people to recall their experiences [47].

Continuous in time. In social science, it is usually hard to keep track of the same research subject, and successfully conduct following surveys or interviews due to many reasons [48]. However, if a user keeps using digital devices or applications, the digital traces are always-on. Digital traces enable researchers to observe user behaviors continuously through a long period of time. With the flexible observing windows, we can compare and analyze the effects of certain events on behavior change.

Large Scale. Cell phones and social media have high penetration rates in many countries [26]. For example, in developed countries, such as the United States, Germany, the United Kingdom, the smartphone penetration rates are more than 75% [49]. Smartphones are usually integrated with web applications and GPS technologies, which enable the generation of digital traces with location and communication information. Given its large penetration rates, behaviors revealed from cell phones and social media can reflect human behavior at large scales, including neighborhoods, communities, countries, or even global scales. Although digital traces can be large-scale data, we need to point out that such data might still have bias in representativeness of the whole population, since not everyone has access to the use of digital devices.

2.3 Data-driven Methods for Three Research Areas

Human activities are impacted by many factors in the living environment, including the relatively stable context such as the physical environment and socio-economic development, and abrupt changes such as natural disasters, shocks, and social mobilization. Large-scale digital traces reflect how people interact with their physical environment or built-in environment. From these traces, we can identify the hidden patterns and correlations that are helpful to support scenario planning and evidence-based decision-making [10]. In this dissertation, I explore the use of data-driven approaches over large-scale digital traces to assist decision-making for Smart and Connected Communities in three research areas: socio-economic development, emergency preparedness and response, and migration studies.

2.3.1 Socioeconomic Development

Socioeconomic information is critical since policy makers usually rely on such information for resource allocation and aid efforts. Household surveys are usually conducted every several years to update this information. However, surveys require extensive financial and human resources. In some places where there is financial deficit or political instability, socioeconomic information is usually incomplete or has not been updated for years. For example, the Southern African Country Angola hasn't updated their census data for more than 40 years due to the civil war, until recently it started the first postcolonial census [50].

Human behaviors are related to the change of context [14]. With digital traces,

we can capture population behaviors and characterize the relations between behavior features and the social context with machine learning models. Since digital traces are generated in an automatic way, digital traces can timely show how people behave at different times, which leads to automatic updates of the knowledge about the socioeconomic context, thus potentially improving the data gathered by surveys.

Ultimately, multiple factors are involved in the determination of population behaviors. It is challenging to disentangle the relations between population behavior patterns and the related socioeconomic context. To contribute to this area, I present a case study of inferring regional socioeconomic levels with population behavior patterns that are extracted from cell phone records. The methods and experiment results are presented in Chapter 4.

2.3.2 Emergency Preparedness and Response

A natural disaster is defined as an event that happens at a specific time and space which incurs losses to the members and damage to the physical structure so that the ongoing functions of the society are disrupted [51]. Different from socioeconomic status, which represents an relatively stable index, this question aims at describing the abrupt behavioral changes when people react to disasters.

Digital traces provide timely information for situation awareness in disasters. Disasters are unique in terms of who is affected and how people are impacted by the environmental change. Although we can gain knowledge from previous disasters, there might be new situations where local governments are not well prepared [52].

Digital traces prove to reflect user needs and specific events in a timely manner, therefore are helpful for efficient reactions in disaster relief [53].

Digital traces have the characteristics of 'large scale' and 'continuous in time', and provide individual-level information. Data-driven approaches using digital traces, therefore, have the advantages of scaling up study to the size of the data and gain resolutions on the spatial and temporal dimension [22]. They can be used to explore user behavior patterns under different environmental contexts, and examine population mobility behaviors across different regions. Large-scale data with fine resolutions also enable the detection of small differences in large datasets that might be otherwise ignored by qualitative methods [14].

In this study, I aim at building data-driven models that can efficiently identify local specific issues for disaster response using online communication data. Citizens and local governments communicate through social media during natural disasters. It is challenging to distill disaster-related information or operational insights out of massive online communications. In Chapter 5, I present a semi-automatic data-driven framework to distill disaster-related issues discussed by citizens. With the identified issues, we are able to reveal local specific issues in targeted areas.

2.3.3 Migration Studies

Migration studies have been conducted in two ways: macro-level and micro-level analyses. Macro-level studies are typically carried out using a combination of various survey and census datasets to model large-scale behaviors, however, these

models fail to provide more nuanced information about the physical or social status of the migrants. Micro approaches, which successfully use interviews and diaries to provide a window into more individual behaviors, could benefit from methods to identify novel or under-studied behaviors that should be addressed in the migration research agenda.

Digital traces usually contain individual-level behavioral information for people in a large geographic area. They provide both micro-level and macro-level perspectives for migrant study. In this thesis, I aim to understand the following two questions: (1) Can we use digital traces to characterize migration behaviors? and (2) Can we provide knowledge of physical, social activities of migrants for a better understanding of the micro-level migration experiences with digital traces?

Migration can be seen as an intervention event to one's life trajectory. Previous work used experimental simulations, survey and investigations to understand the status of subjects before and after intervention [54]. With digital traces, it becomes much easier to observe subjects under multiple control and stimulus conditions [47]. Due to the characteristics of 'continuous in time', we can find the subjects that meet certain pre-defined experimental conditions and observe the change to evaluate short-term or long-term effect with digital traces. Researchers usually use surveys or interviews to collect information regarding pre-event status, however, the information relies on respondents' response, who may not be aware of the details or their memory may not be accurate.

where they may not be aware of details. Digital traces are revealed information that is continuous in time. Studies with digital traces can validate or complement

current studies that rely on survey methods.

In Chapter 6, I present a data-driven framework for migration studies. It uses eight months of cell phone data to observe the continuous change of users, part of which are migrants. With the detected migrants, I evaluate the impact of migrations on their life, from the perspectives of social relations and spatial dynamics. This study provides analyses of migrants' behaviors with angles that are hard to obtain with survey methods.

2.4 Challenges of Developing Data-driven Methods

Data-driven methods with large-scale digital traces can complement current survey methods to collect evidence for decision-making. However, there are many challenges presented in the design of data-driven methods. The key problem is how to re-purpose the existing data and distill valuable and robust information from the data.

2.4.1 Identification of Research Subjects

In the processes that generate digital traces, there might be erroneous data due to the instability of sensors that record the signals of the actions, the temporary malfunction of the system that transfers the data, or errors in the process of saving the data. Social media companies or telecommunication companies take measures to maintain the stability of the services, making sure few errors are generated. However, they are still inevitable. Moreover, sensors may not reflect actual user

behaviors with 100% accuracy. For example, the GPS receiver in a cell phone is usually used to track the trajectories of users movement. However, in mountain terrain and urban areas with high buildings, the accuracy of GPS is usually not high [55]. Multipathing errors might affect modeling of user behaviors, therefore should be identified and removed as much as possible.

Another challenge is to identify research subjects from data since digital traces are generated by different groups of users with different intentions. We need to identify a certain group of people or identify a subset of data that are relevant to the research question. For example, identifying migrants in an unlabeled dataset requires the development of methods to identify migrants based on their spatio-temporal trajectories, which is not trivial. In this dissertation, I will describe methods that I have developed to account for noise in the data and for the adequate identification of research subjects in a large-scale dataset of digital traces that characterize human behavior.

On social media sites, users communicate a wide variety of topics [56]. It might occur that only a small proportion of the communications is relevant for the research question. For example, imagine we are only interested in identifying weather-related tweets in a large dataset. To identify a subset that is relevant to the research question, researchers usually use keywords, hashtags, geotags of tweets, or location of user profiles to filter data. Different retrieval or pre-processing methods can end up with data of different quality, which affect the subsequent data analysis results. In this dissertation, I develop methods to identify specific context within a large-scale dataset of digital traces. In Chapter 5, I present a tweet selection method

to identify disaster-related tweets out of all geotagged tweets, the majority of which are irrelevant to disasters although these tweets were posted during disasters.

2.4.2 Identification of Behavioral Features

Behavioral features are extracted from or computed based on the digital traces. For example, based on spatial locations in time series, we characterize one's mobility trajectories from multiple perspectives, such as mobility distance and mobility frequency. Some researchers hold the opinion that behavioral features should be directed by certain rules and theories, i.e. the extracted features have explainable meanings in sociology [22]. Others argue that theories are no longer required since complete data is self-descriptive, and that we only need to explore the data as much as possible, and to build the model with best performance [57]. In this dissertation, I take the former approach, and extract behavioral features mostly using social theories as guidelines for the identification of specific behaviors.

2.4.3 Identification of Modeling Approaches

Data-driven methods are applied to solve different types of questions in Smart and Connected Communities. As presented in section 2.3.1, for socioeconomic development, we aim at building models that can indicate the relations between digital behaviors and socioeconomic status for the inference or prediction of socioeconomic context. It can be seen as a regression question, in which the targeted values are continuous socioeconomic values, or a classification question, where the targeted

values are discrete socioeconomic levels.

Machine learning models are usually used to address prediction problems. The models are built on a set of behavioral features that are extracted from digital traces. We apply different discriminative algorithms and adjust parameters so that the models can best learn the relation between behavioral features and targeted values. The commonly used discriminative models include Support Vector Machines (SVM), tree-based models such as eXtreme gradient Boosting (xgBoost), ensemble models such as random forest (RF), Bayesian model and neural networks. Evaluation methods are used to assess the quality of models. For regression problems, R^2 is usually used to evaluate how much variance in targeted value can be explained by the model. R^2 is a value scaled from 0 to 1, with values closer to 1 indicating better model quality. Root mean square error (RMSE) is normally used to evaluate the difference between predicted values and the real values on test data sets. For classification problems, the performance is assessed by recall and precision values for each class. Some classification questions require a balanced performance over all classes, some may require better recall for certain class. The evaluation methods can be adjusted to match with the specific goal of the inference/prediction question.

Although previous studies have explored different combinations of behavior features and machine learning models for the prediction of socioeconomic levels, these models mostly use pre-determined features. For example, to characterize mobility behaviors, mobility distance and radius of gyration are used. However, these features fail to capture information such as the direction of the movement, and when the movements happen. To better model socioeconomic development, we

might incorporate features that characterize population mobility behavior with finer granularity. For prediction of regional socioeconomic levels, previous studies used features that represent regions by aggregating individual behavior features using statistical distribution features such as mean, median and percentiles. This type of aggregation fails to capture information at individual-level, which might affect the quality of models. In Chapter 4, I present methods to address these potential limitations in previous studies.

Another type of question is to understand the massive online communications in an automatic way. It can be seen as a clustering question if trained without supervision. Each cluster represents a category of issues that have relatively high semantic similarity. Previous studies also treated it as a supervised classification problem with partial tweets labeled as training data [58]. Supervised methods usually require repetitive labeling efforts for new types of questions. Topic modeling has been widely used as an unsupervised way to reveal topics of massive tweets. However, traditional topic modeling such as Latent Dirichlet Allocation (LDA) is mostly designed for long documents, where multiple topics present in one document. Tweets are mostly short and concise with one main topic. To build an adaptive model, I show in Chapter 5 a single topic model. The key intuition is that a very short document like a tweet is unlikely to be related to multiple topics; therefore it can be modeled as having all its words generated from a single topic. The method tends to generate topics that have better interpretability [59].

Chapter 3: Literature Review

In this section, I summarize the previous studies in the three areas I focus on. For these areas, first I overview the decision-making questions previous studies have worked on. Then I categorize the data-driven approaches based on the behavior features characterized, or the algorithms they proposed, or the strategies of how data-driven models are built.

3.1 Data-driven Methods for Socioeconomic Development

Digital traces have been widely applied in the inference or prediction of socioeconomic contexts. Large-scale collective behaviors can reflect the state of the complex dynamics of social systems [60]. Given the fine-grained data with a high resolution of spatiotemporal information, the data can be used to model the current status of the social system. I present an overview of applying data-intensive methods for different types of socioeconomic development questions, such as happening of crises, land use, and socioeconomic maps. Then I focus on the methods used for inference of socioeconomic maps.

Digital traces have been used to identify the land use, which is the functional area of a city. The reason behind that is individual mobility is highly regular in

the daily base and usually shaped by the context in these areas [61]. Yuan *et al.* proposed a method to model the functional regions (Points of Interests Maps) of Beijing, with GPS trajectories of taxicabs in Beijing with good fitting [62]. Cranshaw *et al.* identified the dynamic boundary of livelihoods by checked-ins collected from the location-based online social network. Such livelihoods are formed based on a distinct characteristic of life, which is different from traditional municipal boundaries and provides vital information for urban planning [63]. Geotagged social media data have been used as a crowdsourced way of understanding functional areas of cities [64–66]. An example is a study that used geotagged Flickr data to identify the closely connected areas and label these areas with Flickr description [65].

Socioeconomic indexes are individual or household values that are aggregated at different granularity levels, such as census blocks, county, and state levels. In developing countries, there are situations where the government makes decisions on missing or out of date data. For example, for antipoverty programs, they need to target the extreme poor to allocate aid resources. Not every government can afford the cost of surveys to acquire such information [50]. Inexpensive and scalable methods to detect the socioeconomic development are especially in urgent need for policy-making in areas with a deficiency in finance.

Over the past few decades, researchers have started seeking solutions for the approximation of different socioeconomic indexes with data. Xie *et al.* found that high-resolution satellite imagery can be used to infer large-scale socioeconomic indicators such as poverty rate [67]. Moreover, results from Jean *et al.* indicated that a model train in one country can be transferred to another, providing infor-

mation for targeting poverty where no or lack of survey data exist [68]. Antenucci *et al.* used social media data to create real-time job marketing indicators such as Job Loss Index, which has a higher updating frequency compared to the traditional index. Social media data are also used to predict the stock market change [69], and consumer confidence index [70].

3.1.1 Analytic Strategies for Inference of Socioeconomic Development

Previous studies use different analytic strategies for the inference of socioeconomic development. Some studies presented machine learning techniques such as SVM regression and classification, and EM clustering to predict socioeconomic maps based on mobile phone data and landlines [71, 72]. Frias-Martinez *et al.* built multivariate time-series models on cell phone behavioral features and find that consumption and mobility features extracted from call detail records can be used to forecast the socioeconomic time series by the National Statistic Institute (NSI) [73]. Xie *et al.*, on the other hand, used deep learning algorithms to extract socioeconomic indicators from high-resolution satellite imagery and achieve high prediction accuracy with better granularity [67]. njuguna *et al.* combined CDRs and satellite imagery to estimate the socioeconomic development in sector level, which yields competitive results and can be an enhancement to survey methods [74]. Toole *et al.* built a structural break model to identify the unemployment status of cell phone users, and aggregated individual level results to predict province level unemploy-

ment rate, which achieved high correlation with actual unemployment rates [75]. Almaatouq *et al.* built unsupervised and supervised Gaussian Processes (GP) models to predict district level unemployment rates based on the population behavioral characteristics of residents in each district [76].

3.1.2 Behavioral Features Relating to Socioeconomic Development

A set of previous work focused on examining the relations between human behaviors and socioeconomic contexts. Several behavioral features can be potentially used to infer the current socioeconomic status or predict future changes. In the following, I summarize the categories of behavior features that have been used.

Information searching and communication behaviors. Information searching and communication behaviors can reflect the employment status of a society. Researchers have used search query data and social media data to capture the behavioral signals for prediction of socioeconomic status. Antenucci *et al.* used phrases related to job loss, job search, and job posting on social media as signals for the prediction of employment indexes. As opposed to traditional indexes, the index inferred from social media data has higher temporal granularity and are of value to policymakers in need of real-time information [77]. Li *et al.* analyzed geotagged tweets and photos from Flickr, and found features such as tweeting frequency and photon density that are correlated with socioeconomic characteristics such as average income [78]. Meanwhile, certain topics on tweets also tend to be related to community well-being [79].

Emotions. The expressions of emotions are indicators of socioeconomic development. Bollen *et al.* extracted six mood states which are tension, depression, anger, vigor, fatigue, and confusion from the aggregated Twitter data to infer the existing social and economic indexes [44]. They found that the large-scale population mood is highly related to the socioeconomic sphere. Large-scale sentiment analysis is also proved to be useful for the prediction of consumer confidence index [70].

Cell phone use behaviors. Cell phone use behaviors are a combined set of behaviors including mobility, social ties, social activities, and consumption features that are extracted from cell phone records. Previous studies have examined different combinations of cell phone use behaviors as predetermined features for the inference of socioeconomic status. Empirical studies show that there is a high correlation between individuals' communication behaviors and their household income [80], and other welfare indicators such as assets, housing, and health [81]. At the region level, Smith *et al.* extracted a set of call behavior and mobility features from cell phone data and explore a linear regression model to approximate poverty level [82]. Smith *et al.* used social interaction features from cell phone data to identify the regional poverty level of a developing country [83]. Soto *et al.* presented methods that explored features such as activity range, communication reciprocity to predict the socioeconomic levels defined by the National Statistical Institute [84]. Eagle *et al.* examined the structure of social networks and the economic development of communities and found that the diversity of individuals' relationships is highly indicative of the local economic development [72].

Socioeconomic levels are also related to other types of signals. Gutierrez *et*

al. combined the history of airtime credit purchases and communication records to estimate the relative income of individuals and the diversity and inequality of income [85]. The brightness of light at night also proves to be highly correlated with an asset-based wealth index in Africa [86]. Jean *et al.* demonstrated a computing method for estimation consumption expenditure and asset wealth using high-resolution satellite imagery, which helps to support target aid efforts [68].

3.2 Data-driven Methods for Emergency Preparedness and Responses

3.2.1 Digital Traces Used in Different Phases of Disasters

People use social media to communicate emergency situations during natural disasters [87], connecting with friends and families and share information such as locations medical services and shelters [88]. Users also show altruistic actions to voluntarily contribute geotagged information describing the flooding phenomena [89], or provide social support by spreading information about missing people, raising funding, and offering necessities and shelters [90]. Social media sites become critical platforms for people to communicate. This also motivates governments and organizations to interact with their residents online [91]. They gain valuable information for situation awareness, to address challenges in different phases of disaster prevention, preparation, response, mitigation, management, and recovery [92].

Some focused on a specific phase of disasters. For example Carley *et al.* examined the use of social media data for early warning and preparedness in disasters by identifying the opinion leaders for information dissemination [93]. Bruns and Liang

introduced different approaches of analyzing Twitter data for disaster response [94]. Some targeted using social media data for all stages of disaster relief [95]. Due to the richness and complicity of social media data, Horita *et al.* proposed a framework that helps to align the decision-making problems in organizations with data sources that help to solve these problems [96]. Digital traces have been used in the following categories of studies.

Displacement of people in natural disasters. Disasters usually cause damage to the environment, infrastructures and affect the routine life of residents. A certain level of disruption can cause people to move away and seek safe places to settle [97]. Understanding and predicting of population mobility behaviors are critical for government and organizations to plan, support and respond to disaster evacuations. Researchers have tried to model mobility behaviors in different ways. Hara and Kuwahara used smartphone GPS data to understand the evacuation behaviors, mostly focusing on road network conditions. The data showed the gridlock phenomenon after the disaster, which is serious traffic congestion in the central area of the city [98]. Wang and Taylor used geotagged tweets to quantify human mobility distances before, during and after disasters. They found there are variations in short and long trips during disasters, although mobility patterns still follow the Levy-Walk model [99].

Communication behaviors during natural disasters. Online communities have significantly contributed to the disaster response, recovery and mitigation [100]. Governmental agencies need to understand community level behaviors for better relief policies. The interactions between users on Twitter have been stud-

ied a lot, for example, in reply to and retweet activities [101–103], or social network features [104]. These work mostly focused on the network analysis of users’ social activities. Hughes *et al.* analyzed the citizens @ behaviors when reporting issues to the local government accounts in general [105]. Eriksson *et al.* showed a comparative analysis of citizens and crisis communication professionals, focusing on the perceived usefulness of content [106].

Communication contents of social media data. Researchers used historical social media data to understand user needs during natural disasters. Kaigo showed that geotagged tweets can be used to locate the needs of electricity, gas, food and other essential items [87]. Kenneth *et al.* found that social media content is valuable for identifying specific needs and concerns of local residents [53].

Sub-event detection with social media can potentially increase the local situational awareness and therefore improve disaster management. Chae *et al.* designed an interactive spatiotemporal visualization tool that assists decision-makers by identifying the trending local events discussed in social media platforms [107]. Pohl *et al.* proposed a clustering method for sub-event detection with geotagged tweets. Others categorized the type of communication information [108, 109], or used tweets as a proxy for damage assessment during disasters [110].

3.2.2 Understanding Online Communications for Disaster Response

Governments and disaster relief organizations need to understand the stream of messages online so that they can monitor the public safety-related issues and make

efficient responses. However, it is challenging since the data volume is enormous, while contains substantial noise [111]. Both supervised and unsupervised methods have been used for the understanding of online communications.

Supervised Methods Supervised methods usually incorporated manually labeling work to summarize discursive themes in communication during natural disasters [112], to extract valuable pieces of information [58], and to explore the common reaction patterns of users across different types of disasters [113].

Some of the studies aimed at building a list of crisis-related terms across different types of disasters. These crisis dictionaries can be further used to retrieve disaster-related information for situational awareness in other cases [108]. The process of curating crisis-related terms, however, requires manually reviewing a large number of tweets. Due to the limitation in time, only a small proportion of sampling tweets are labeled [108]. Olteanu *et al.* also proposed an analytical framework to explore the common user reaction patterns to different types of disasters. They analyzed crisis tweets from the perspectives of crisis type and content type. It is also based on the labeling of sampled tweets [113]. Starbird and Palen examined retweet behaviors with the content of tweets, which was acquired by manually reviewing and labeling [114]. These types of study are valuable in that they demonstrated the common keywords used, the knowledge of how user react to different types of disasters and what kind of valuable information can be extracted. However, both dictionaries and common pattern knowledge have generality but lack specificity. It is still in need of efforts for identification of local specific issues in a certain disaster scenario.

There were cases of studies analyzing tweets during several disasters. Acar and Muraki have studied crisis communication during Japan’s tsunami disaster. They looked into tweets of three categories: warnings, help requests, and reports about the environment and self [115]. Shaw *et al.* chose tweets during the floods in Queensland for disaster communication study. They categorized particular genres of tweets based on the purposes of communication: direct information, media sharing, help and fundraising, personal experience and discussion. Similarly, they sampled 5% of the tweets and label tweets by categories. However, the boundary between different categories seems not to be crystal clear [112]. In the study on tweets during the landing of hurricane Sandy, Spence *et al.* also used pre-determined categories to understand the crisis communication online. The categories they set include *Information*, *Affect Display*, *Humor*, *Insult*, and *Spam*, which are different from previous studies [116].

Another line of research involving human labeling work is to train computational linguistic models with labeled training data. Imran *et al.* defined categories of personal, informative and other irrelevant and built a conditional random field (CRF) model to automatically separate tweets into these three classes [58]. The model trained on one disaster can be used to categorize tweets of other disasters. However, it also has the problem that the classification provides little information that is specified in a certain disaster.

Unsupervised Methods. Besides manual coding, previous studies have used unsupervised methods to cluster topics and to explore potentially valuable information out of tweets [117]. MacEachren *et al.* presented a map-based visual analytic

system that used the explicit and implicit geographic information and contents of tweets to build the place, time and theme index. Spatiotemporal analysis on the top of the index enables to find critical issues during crisis or disasters [118]. Chae *et al.* presented an interactive visual analytic system, where users can examine the topics when there is an abnormal change of tweet volume in temporal trends [9]. Chae *et al.* also proposed a visual system with spatiotemporal analysis integrated. This type of visualization systems can support situation awareness for in-time reaction during natural disasters [107].

However, they work as assistant tools and require experts to interact with these systems to explore potentially valuable information. Spatiotemporal bursts detection with tweets posted during disasters does not necessarily reveal all disaster-related issues the local government should react. During natural disasters, more users tweet words, scenes, news about disasters [89]. But the majority of users tweet about other topics irrelevant [59]. It is highly probably that disaster-related topics may draw attention and present as a spatiotemporal burst in the timeline, but there can be burst topics irrelevant to disaster. While some disaster-related issues that should be paid attention can be ignored due to not fitting to burst pattern.

Rule-based models have been used to detect the key elements regarding public safety issues. Xu *et al.* proposed a 5W (What, Where, When, Who, Why) framework to understand social media posts during emergency events. Each element is automatically extracted from a post with entity recognition rules [119]. This method is useful to understand details of a certain event described in each tweet, but not very helpful to discover categorized issues at a large scale.

3.3 Data-driven Methods for Migration Study

There are two types of analyses for the study on migration behaviors: macro-level and micro-level analyses. Macro-level studies are typically carried out using a combination of various survey and census datasets, including origin-destination internal flows as well as demographic and socioeconomic data, to assess the role those specific variables might play as both determinants or consequences of migration movements [120]. These macro-level studies provide general migration trends that are highly useful from a policy perspective. Decision makers can assess the types of social groups, characterized for example by age or profession, that are migrating between regions, and the long term impact these migrations have, for example, on the local economy. However, the macro-level analyses fail to evaluate more nuanced variables that can be captured through micro-level analyses including data from interviews and diaries [121].

Micro-level analyses provide a window into the physical, social and psychological status of internal migrants showing, for example, that migrants maintain strong social ties with the communities they leave behind [122]; that migrants show different spatial behaviors to locals, mostly due to search processes in the physical environment that generate spatial dynamics that differ from those that locals show [123]; or that internal migrants usually encounter difficulties in adapting to new environments, suffering from a series of issues such as psychological stress which might affect their behavior in the physical environment [124–126]. Although the migration research agenda at the micro-level is broad, most of these studies chose

a sample area to conduct surveys, lacking evidence whether the reveal phenomena applied to other social contexts [127].

The following section summarizes previous studies for macro-level and micro-level analyses with crowdsourced data or survey methods.

3.3.1 Migration Models Built on Crowdsourced Data

Limited data availability has been one of the main bottlenecks for empirical analysis and theoretical advances in the study of migrations [32]. Since aggregated data resolved in time and space can describe the large-scale collective behavior of users [33, 128, 129], it has recently been used for many studies of migration.

Zagheni *et al.* used email service logs to identify international migration rates [130]. Hawelka *et al.* used geotagged tweets as a proxy to model global mobility patterns [131]. Weber *et al.* used anonymized users' log into Yahoo! services to generate short-term, medium-term mobility flows across countries. Considering the geographic accuracy of the IP address, they focus on international migration rates [132]. The geotagged tweets can reveal users' location in the scale of a city or even a specific point. It has been used to improve understanding in the internal migration flow since census data may have inconsistency over time and inaccurate due to lack of survey samples [32].

Nevertheless, all these approaches suffer from a large bias problem, since the demographic and economic backgrounds of email, web and Twitter users are not representative of the population at large [133]. On the other hand, cell phone data,

with much higher penetration rates across all types of people, has been shown to be more representative, although still imperfect, of the population at large [26]. As a result, Blumenstock *et al.* proposed a macro-level method that used cell phone metadata to identify migrants and quantify volumes and directions of internal migrations in Rwanda [134]. All these studies mainly focused on the macro-level analyses of examining the migration flow of large population.

3.3.2 Survey-based Migration Analyses

Survey-based studies on internal migration have focused on the use of macro- and micro-approaches to assess pre- and post-migration behaviors brought up by migratory movements [135, 136]. The features used in macro approaches typically differ from the micro approaches due to the nature of the datasets they use: while macro approaches typically focus on large-scale census data, micro approaches are prone to use survey and interview data at smaller population scales.

Macro-approaches have been used to understand the factors that drive migration flow. Previous studies have shown, using census data, that internal migrants tend to relocate due to unemployment, lack of services, poverty or lack of safety to areas that offer better conditions [137, 138]. These features have been used by researchers, to build theoretical models that explain and predict migrations at the macro scale, including Zipf’s inverse distance law [139], Stouffer’s law of intervening distances [140] or Gravity models [141].

On the other hand, micro approaches have mostly focused on analyzing be-

havioral features that characterize the mobility, social activities or psychology of migrants. The studies with behavioral features are numerous, including Li *et al.* which showed that migrants in Beijing have restrained spatial dynamics which are heavily skewed towards people from the same town of origin [142]; Nogle *et al.* that analyzed the importance of offering migrants settlement assistance as well as information about local opportunities [143]; Hendriks *et al.* observed that migrants tend to spend less time in social activities [144]; Gurak *et al.* who showed that migrants' life status can be affected by factors such as kin and friend relationship [122]; Kuo *et al.* who related the adaptation problems of migrants to the difficulties in the restoration of disrupted social networks [145]; and Chib *et al.* who showed that migrants rely on social support (emotional, instrumental or information aid) to deal with the stress caused by their migration experiences [146].

Some studies focused on the demonstration of psychological changes in the attempt of re-establish lives in the host society [124, 147]. Mou *et al.* showed that migrants usually have a higher level of stress, associated with smoking, frequent Internet usage and long working hours [148]; or are unhappier than locals and spend less time in happiness producing social activities [144].

Chapter 4: Study 1: Inferring Socioeconomic Maps with Topic Models on Mobility Motifs

Socioeconomic maps are important to policymakers for understanding regional development. While computing these maps is oftentimes time-consuming and costly, crowding source data such as cell phone records contain mobility information of users, from which we may extract socioeconomic characteristics. Such characteristics can be used to infer regional socioeconomic levels. I propose four methods of inference based on: 1) the pre-determined human behavioral features; 2) spatiotemporal transition features; 3) the population mobility patterns extracted from spatiotemporal transitions by the topic model Latent Dirichlet Allocation (LDA); 4) supervised LDA on spatiotemporal transitions. I evaluate prediction accuracy for the four different methods and find that the topic model can help extract patterns of co-occurring mobility behavior features and improve prediction accuracy. Specifically, the supervised topic method (sLDA) achieves the best performance and improves the state-of-the-art prediction by 9%. Topic models have been used in various scenarios other than natural language processing, such as image processing [149], spatiotemporal data [62] and social networks [150].

4.1 Research Question

Socioeconomic maps gather information relating to or involving a combination of social and economic factors, such as household income or poverty rate, in various geographic scales. It reflects the socioeconomic development of certain regions, which is important information to policy decisions made by the government and international organizations. For example, the World Bank, which aims to eliminate poverty, heavily depends on poverty incidence data for distributing loans and offering advice.

Socioeconomic information is usually gathered by interviews or investigations, aggregated and reported at various granularity levels, from metropolitan areas to cities, states, or even worldwide countries. Such information has to be updated every several years to reflect the changes in socioeconomic development. The computation is usually highly expensive because it involves a large number of people to carry out interviews and to collect and analyze data. On the other hand, the ubiquitous use of cell phones provides an indirect way to understand human behaviors, which helps to infer socioeconomic status. Cell phones are widely used in modern societies, even in some rural areas [80]. Cell phone data can provide a rough picture of how humans move, showing their interaction patterns across time and space. Meanwhile, human mobility follows some regular laws relevant to socioeconomic levels [7].

Previous studies have worked on the relationship between socioeconomic status and human cell phone behaviors on both micro (individual) and macro (regional) levels, which proves the possibility of SEL inference based on cell phone data. From

the personal perspective, empirical studies show that cell phone based behaviors, such as social network and travel distance, are correlated to specific individual socioeconomic characteristics [151]. From the regional perspective, Newman *et al.* and Eagle *et al.* find that the overall communication behaviors of a region could be highly correlated with the socioeconomic development [72, 152]. Researchers have used machine learning approaches to predict regional socioeconomic levels on the overall usage of cell phones in each region [153].

I further explore a new method of inferring socioeconomic levels (SEL) from cell phone behaviors at the macro level. Specifically, the concept of mobility motifs is defined to represent spatiotemporal transition status. It uses topic models to extract population mobility patterns from mobility motifs as features for SEL inference. The method achieves better performance than state-of-the-art prediction algorithms, which improves accuracy by 9%.

4.2 The Proposed Method

Figure 4.1 shows the methods explored in this study. Each geographic unit is represented by a pair of SEL associated with the spatiotemporal data recorded by the tower in this region. The SEL can be a continuous variable, such as the poverty rate, or a discrete value indicating the class. Depending on the SEL, the problem can be treated as either regression or classification.

The four methods are represented as Pre-determined Features (PF), Pre-determined Features II (PF2), extraction of Population Mobility Behaviors by La-

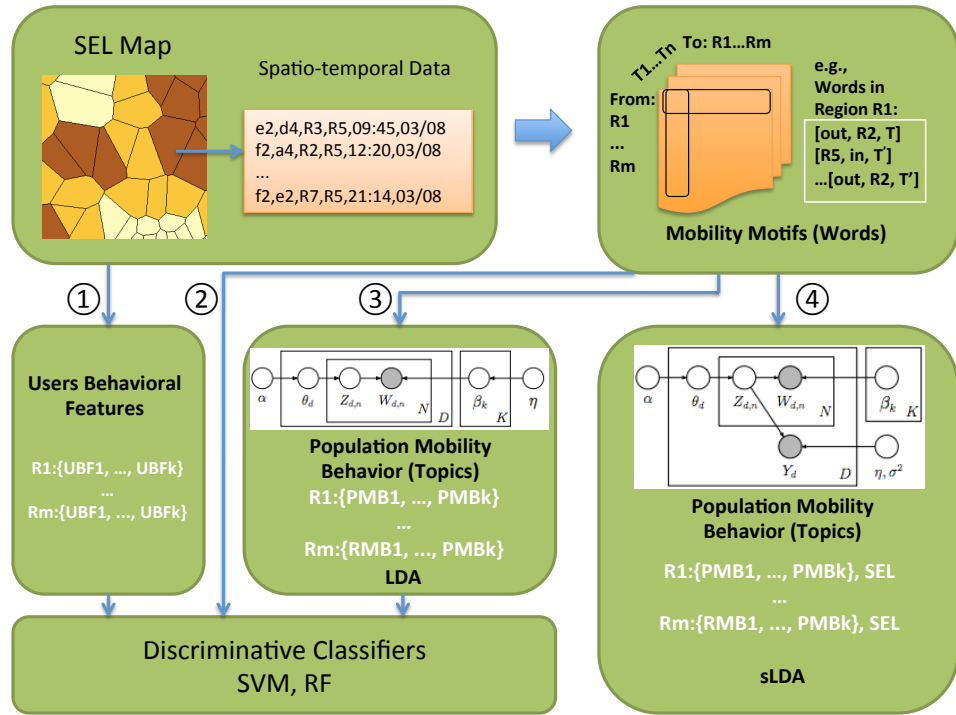


Figure 4.1: Approach overview: ①PF ②PF2 ③PMB-LDA ④PMB-sLDA. LDA and sLDA plate notation from [1]

tent Dirichlet Allocation (PMB-LDA), and Supervised LDA on Population Mobility Behaviors (PMB-sLDA).

4.2.1 PF

In this approach, a vector containing pre-determined behavioral features, which are designed on the hypothesis about mobility and socioeconomic level, represent each region R_i . I present two groups of features: user-based and tower-based features. User-based features characterize individual behaviors, and tower-based fea-

tures characterize general social (group-based) features associated with the geographical area covered by each tower.

The SEL of each region is determined based on the individuals who live there. The hypothesis is that any calling activity after 6 pm will highly probable to be associated with the home location. For users who have only one such place, the tower is assigned as the home tower. If an individual has more than one tower associated with her activity, home is determined as the tower which percentage of activity after 6 pm represents at least 30% of the overall activity for that user. If such tower does not exist, the home location cannot be identified. Thus the user's behavior won't be taken into account to compute the user-based features. However, it is considered for the tower-based features.

The following user-based features are defined: consumption features, mobility distance, the radius of gyration, mobility entropy. Tower-based features represent aggregated activities at a tower level. These features measure all call activities that have been taking place in the area covered by the Voronoi polygon that approximates the coverage of the cellular tower.

Other than mobility entropy, the distributions of user-based features are heavily skewed. The regional feature value cannot be simply computed by mean of individual values since a small proportion of large values can highly affect the average value of a region. For these features, I calculate the composition of low-, medium-, and high- subgroups to characterize each region. Specifically, based on the distribution of feature values of all individuals, two thresholds are identified to define three ranges. For example, for the consumption feature, individuals with less

than 28 output calls belong to the low output subgroup; more than 66 belong to the high; the rest belong to the medium. For each region, I calculate the percentage of individuals fall in the three ranges. The threshold values are set based on correlation analysis between percentage values and SEL. At the end of this process, each tower region is characterized by a set of features based on user-based behaviors. Each feature is represented by a set of three ranges of characterizing values per sub-group.

Tower-based features represent aggregated activities at a tower level. These features measure all call activities that have been taking place in the area covered by the Voronoi polygon that approximates the coverage of the cellular tower. The following tower-based features are defined: Newcomers, new visits, the proportion of newcomers to residents, the number of input/output calls.

After the representation of region features, a correlation analysis is used to detect features that are highly related to SEL. Then, regression and classification methods are applied to train models based on these features. The algorithm is

described in Algorithm 1.

Algorithm 1: PF

```
1 foreach Region  $R_i$  do
2   foreach User feature  $F_j$  defined in  $R_i$  do
3     Calculate  $F_j$ ;
4     Update  $R_i = [F_1, \dots, F_j]$ ;
5   end
6    $RS_i = (R_i, SEL)$ ;
7 end
8 if  $SEL \in \mathfrak{R}$  then
9   Train and test on  $\cup_{i=1}^N RS_i$  using SVR and RFR;
10 else
11   Train and test on  $\cup_{i=1}^N RS_i$  using SVM and RF;
12 end
```

4.2.2 PF2

In the method of PF, the features are built based on the hypotheses between socioeconomic development and human behaviors. However, the features are composed with statistic aggregation of individual behaviors, which may lose details that can be important for discriminating regional SEL, for example, the information of when and where to go in mobility features.

PF2 method directly uses individual mobility activity as features for regions. Mobility motifs are defined as transition activity containing origin and destination

regions associated with the time range when the event happens. As a result, each region is represented by such motifs as features. Specifically, the mobility motifs are derived from CDRs as follows. Suppose one cell phone record is represented as (i, j, R_i, R_j, T, D) , where a phone number i in region R_i calls to phone number j in R_j at time T and day D . Then, two consecutive records from one person generate one mobility motif. For phone number i , if the next record is $(i, k, R'_i, R_k, T', D')$ where a combination of $(T', D') > (T, D)$, it indicates the activity of i from R_i to R'_i . Then, we can extract a mobility motif for region R_i as (out, R'_i, T) , which means an individual outgoing transition to R'_i at time T . At the same time, R'_i has an involved transition (in, R_i, T') , which means an individual incoming transition from R_i at time T' . A call record means the status of someone in some region, but not transition status. A precise time point does not mean the time of leaving or arriving. So, the time is discretized into six four-hour ranges, i.e., $T \in \{[0, 4), [4, 8), \dots, [20 - 24)\}$. On the one hand, it avoids the error caused by the precise timestamp; on the other hand, it reflects the population mobility trend. For example, a trip in time range $[16, 20)$ is probably going home from work.

By this way, each region is represented by the mobility motifs related. The discriminative algorithms described in [4.2.1](#) are applied.

4.2.3 PMB-LDA

In the method of PF2, mobility motifs are features. However, the variety of mobility activity leads to excessive data dimensions, which requires high computa-

tion consumption, and some spatiotemporal activity features may not be correlated with SEL. In this part, I propose a generative method to extract large probabilistic behaviors, which can be interpreted as population mobility behavior (PMB). Mobility motifs in geographic units can be used to discriminate regional SEL since these motifs reveal functions of regions that are related to SEL. The function of a region causes a certain pattern of mobility. Under this scenario, I firstly use an unsupervised method to reveal PMBs across regions and use the PMBs as features to infer regional SEL.

LDA is widely used for detecting topics of documents; thus, the meaningful co-occurrence patterns of words. To find spatiotemporal patterns out of mobility motifs, I apply LDA to extract the PMBs (topics) from mobility motifs (words) of regions. Each region is composed of a distribution of topics with different proportions, while each topic is composed of a distribution of words. The topics are shared across all regions. Each region is a combination of topics with different proportions. A vector of topic distribution is used as input to discriminative regression and classification algorithms to infer the SELs.

The LDA requires the number of topics as a preset value. However, it is hard to identify how many topics are contained in the spatiotemporal data, so I execute multiple experiments and identify the number of topics when the prediction accuracy

is highest. The process is shown in Algorithm 2.

Algorithm 2: PMB-LDA

```

1 Draw PMB proportions  $\theta|\alpha \sim \text{Dirichlet}(\alpha)$ ;

2 foreach mobility motif do
3   | Draw PMB assignment  $z_n|\theta \sim \text{Multinomial}(\theta)$ ;
4   | Draw mobility motif  $w_n|z_n, \beta_{1:K} \sim \text{Multinomial}(\beta_{z_n})$ ;

5 end

6 foreach region do
7   |  $R_i = [PMB_1, \dots, PMB_n]$  (n is number of topics)

8 end

9 if  $SEL \in \mathfrak{R}$  then
10  | Train and test on  $\cup_{i=1}^N RS_i$  using SVR and RFR;

11 else
12  | Train and test on  $\cup_{i=1}^N RS_i$  using SVM and RF;

13 end

```

4.2.4 PMB-sLDA

Other than the unsupervised way to generate PMBs, they can also be generated under the known SEL values, which means the set of latent topics or PMBs are generated in a supervised way. I use the supervised Latent Dirichlet Allocation

(sLDA) to implement this, which is described in Algorithm 3.

Algorithm 3: PMBSEL-sLDA

```

1 Draw PMB proportions  $\theta|\alpha \sim \text{Dirichilet}(\alpha)$ ;

2 foreach mobility motif do
3   | Draw PMB assignment  $z_n|\theta \sim \text{Multinomial}(\theta)$ ;
4   | Draw mobility motif  $w_n|z_n, \beta_{1:K} \sim \text{Multinomial}(\beta_{z_n})$ ;

5 end

6 if  $SEL \in \mathfrak{R}$  then
7   | Draw SEL  $y|z_{1:N}, \eta, \theta^2 \sim N(\eta^T \bar{z}, \theta^2)$ ;
8 else
9   | Draw SEL  $y|z_{1:N} \sim \text{softmax}(\bar{z}, \eta)$ 

10 end

```

\bar{z} is defined as the empirical frequencies of PMBs in the region. The response comes from a linear model, while η indicates the regression coefficients. In this process, the document is generated first, thus sampling process of words and topic assignments. Then, based on the document, the responsible variable is generated. The regression coefficients directly join in the optimization process. In contrast to the PMB-LDA model, where the regress variable depends on the topics distribution and has nothing to do with the topic generation process, this method generates topics that can better explain the variance in the response variable. It is proved to improve predictive performance in [154].

4.3 Results

4.3.1 Dataset

The data used for experiments is Call Detail Records (CDRs). In cell phone networks, base transceiver stations (BTS) or cellular towers are in charge of giving coverage to cell phone devices. Each area covered by a BTS tower is called a cell. For simplicity, the cell of each tower is a two-dimensional and non-overlapping region segmented by Voronoi diagrams. It is a widely used method in geography to segment spaces based on the location of points [26]. Whenever an individual makes a phone call in a cell, the call is transmitted through the tower and generate a CDR. The CDR contains information of encrypted phone number, call time, type (in/output call), duration and the tower that transmit signals. It offers information about who stayed in which region at that time. A time series CDR for one person reflects his rough mobility trajectory. After removing super numbers that have an excessive number of calls or involve with too many cell tower from the data, there are 134 million calls by 1.8 million individuals. The records with these numbers are removed since they tend to be business or spam numbers that cannot reflect individual behaviors.

In this study, the SEL is municipal poverty incidence provided by the World Bank. It represents the proportion of the population below the poverty line in the municipal area. The SEL values are continuous. For completeness, I also separate SEL values into three different ranges and report the result of classification. The classification results can determine whether the method accurately predicts the so-

cioeconomic level one region falls in. Finally, there are 48 regions labeled as level A (lower poverty incidence, high socio-economic level), which represents high SEL, 78 regions with level B and 67 regions with level C (high poverty incidence, low socio-economic level).

4.3.2 SEL Inference

With the above data, each geographic unit is represented by the feature vector and the SEL. I frame it as a regression or classification problem based on the SEL values.

To test whether the features are indicative of SEL, I randomly divide the 186 units into a 75% part as the training dataset, and 25% as the testing dataset. Each time, the training data are used to train the support vector machine (SVM) or Random Forest (RF) model, the inference accuracy on test data are reported. To ensure the model is biased because of the data division, I repeat this process for 100 times and report the average accuracy across all runs. For regression problem on continuous SEL values, R^2 and RMSE are used. For discrete SEL values, I calculated precision and recall per class and reported F1 score, which is a value reflecting the combination of precision and recall. Meanwhile, because accuracy for each class may be biased toward the majority, I also report average F1 score and average accuracy to evaluate the overall performance $F1 = \frac{2*Precision*Recall}{Precision+Recall}$.

Similarly, the discriminative models are applied to the mobility motif features described in 4.2.2. In total, there are 4.4 million mobility motifs across all regions,

with an average of about 23,700 motifs per region. Using the motifs extracted in PF2, I conduct a series of experiments by adjusting the number of topics for PMB-LDA and PMBSEL-sLDA. As Figure 4.2 shows, PMBSEL-sLDA performs best when we set the number of topics as 25. For PMB-LDA, either with algorithm SVM or RF, 20 topics have the best accuracy.

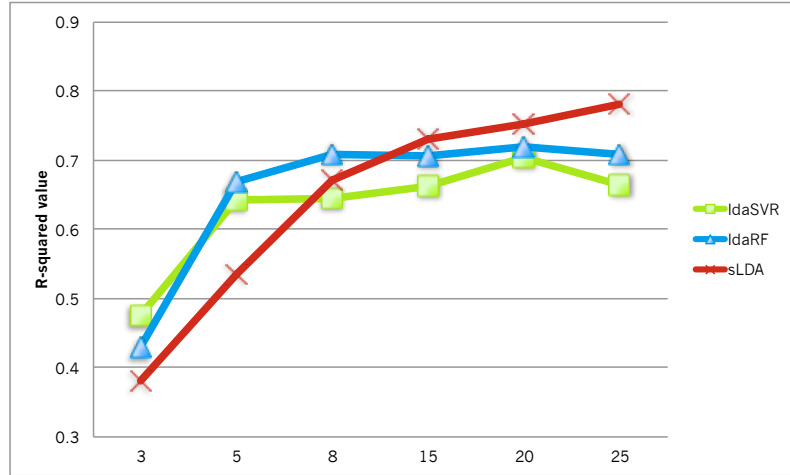


Figure 4.2: R^2 per number of topics for PMB-LDA (SVR and RF) and PMBSEL-sLDA approaches.

The regression results are shown in Table 4.1. Both topic model approaches have the relatively better R^2 and lower RMSE. In PMBSEL-sLDA, the R^2 is highest as 0,7802, which means the topic features can explain about 78.02% of the total variance in SEL. Meanwhile, the R^2 increases 0.1514 compared to the RF on user behavior features, and 0.0875 in PF2, which is the best case in pre-determined features. In comparison, the supervised topic model on mobility motifs (PMBSEL-sLDA) is better than the methods on-topic features by the unsupervised topic model (PMB-LDA). The R^2 increases about 9%, and the RMSE decreases about 15%, meaning

that supervised topic models can reveal latent population mobility behaviors (PMB) that are highly indicative of SEL. This result is consistent with another study conducted by [1]. Also, the result shows that PMB is better used for inference of SEL than simple mobility motifs even they contain all activity information about regions, directions and time. It is probable that in mobility motifs there are activities that are not highly related to regional SEL and can be noise if taking into account. Topic model in a way reduces feature dimensions and extracts the principal components that are related to SEL.

| REGRESSION | | R^2 | RMSE |
|-------------|-----|---------------|--------|
| PMBSEL-sLDA | | 0.7802 | 0.0902 |
| PMB-LDA | SVR | 0.7050 | 0.1088 |
| | RF | 0.7188 | 0.1058 |
| PF | SVR | 0.2573 | 0.1731 |
| | RF | 0.6927 | 0.1156 |
| PF2 | SVR | 0.5721 | 0.1290 |
| | RF | 0.6288 | 0.1195 |

Table 4.1: Accuracies for regression with topic models and pre-determined features.

The classification models built on SEL classes also show similar results (Table 4.2). Both topic models show an improvement in performance, including average accuracy and average F1 score. The supervised topic method (PMBSEL-sLDA) increases about 6% for best case in pre-determined feature methods and about 4% for PMB-LDA. PMBSEL-sLDA achieves relatively balanced results across all classes. Although RF in PF2 has a good average F1 score, the F1score is biased in the three classes. The methods show spatiotemporal data are highly correlated with SEL. Through topic models, we can extract population mobility behaviors, which can explain much variance in the regional SELs.

| CLASSIFICATION | | ACC | AVG.F1 | F1 | | |
|----------------|-----|--------|---------------|--------|--------|--------|
| | | | | A | B | C |
| PMBSEL-sLDA | | 0.7565 | 0.7526 | 0.7273 | 0.7283 | 0.8023 |
| PMB-LDA | SVM | 0.6237 | 0.6302 | 0.6609 | 0.5519 | 0.6777 |
| | RF | 0.7130 | 0.7212 | 0.7786 | 0.6572 | 0.7276 |
| PF | SVM | 0.6200 | 0.6374 | 0.7409 | 0.5586 | 0.6128 |
| | RF | 0.6440 | 0.6567 | 0.7468 | 0.5847 | 0.6387 |
| PF2 | SVM | 0.4522 | 0.4510 | 0.4195 | 0.4198 | 0.5139 |
| | RF | 0.7004 | 0.7100 | 0.7856 | 0.6283 | 0.7160 |

Table 4.2: Accuracy(ACC), average F1 and per-class F1 score with topic models and pre-determined features.

4.4 Discussions

This study explores different methods of inferring SEL maps by cell phone data. Specifically, it shows that topic models on spatiotemporal data can enhance the existing approaches on predetermined features. The method of PF works mainly from a micro perspective: the distance and mobility variety of users. However, cities may have different urban form and radius, which also affects the commuting distance and distance for accessibility. Special case, for example, is that rich people in cities with a large radius would prefer living in the center of a city, where their travel distance won't be high. They can also have high accessibility to opportunities and services in the Central Business District (CBD); such activities would not contribute to the variety of mobility. While, public transportation enables people to travel far with a bearable cost. So, low SEL people may spend more time traveling to minimize the necessary cost of housing. The features in PF depend on assumptions of similar urban form and use of public transportation, which, in reality, differs a lot in cities. Although PF in our situation achieves acceptable accuracy, the inference

performance may vary for different cities or countries.

In consideration of that, PF2, PMB-LDA, and PMBSEL-sLDA are more robust. The methods only use the spatiotemporal data, which are mobility transition statuses across different regions. Mobility motifs only take into account the place and time and do not include any information about distance. Meanwhile, topic modeling methods also provide important information about which topics are related to socioeconomic development.

4.5 Limitations

Data-driven methods have limitations which come from biased representation of data. First, the penetration rate of cell phone use is not 100% in the country I studied. There are groups of people, especially those who have socioeconomic disadvantages in certain geographic regions do not have access. Therefore, data about these people are unavailable. They are the group of people who need help and should be considered in the socioeconomic development plan. In this situation, cell phone data cannot substitute the role of human investigation. The bias caused by the data should be considered.

Second, the digital trajectories extracted from CDRs are only approximation of real mobility traces. In general, people who travel a longer distance show longer mobility distances in digital trace. However, mobility that happen in small scopes (e.g. inside the coverage area of one tower) are not reflected in CDRs. If people move without frequent connection to cellular towers, their mobility is also not reflected in

digital traces.

Chapter 5: Study 2: Understanding Online Communications during Natural Disasters

A growing number of citizens and local governments are using tweets to communicate during natural disasters. A good understanding of the communication contents and behaviors is critical for disaster relief. Previous work has used crisis taxonomies or manually labeling methods to understand the content. However, such methods usually require extra efforts to find insights related to specific events. In this study, I use a semi-automatic framework to extract topics from the communication contents of citizens and local governments, combined with the spatiotemporal information to explore: 1) the spatiotemporal bursts of topics; 2) the change of topics with respect to the severity of disaster; 3) communication behaviors. I use tweets collected during 18 snowstorms in the State of Maryland, US. The study reveals user needs in different communities.

5.1 Research Question

There have been a growing number of people using social media for communication during natural disasters such as snowstorms or floods [155]. Twitter has been used by organizations or local governments as an information-spreading tool [114];

as a platform for situational awareness [156]; or to connect and engage with citizens [155]. However, local government or organizations usually lack the information, policies or guidelines to make communication strategies as when to post what kind of information in emergencies [157].

A good understanding of citizens online communications during natural disasters would allow local governments to better cater to their needs [158]. However, identifying the specific issues for certain disasters is not straightforward. There have been attempts by utilizing taxonomy with predefined sets of crisis communication categories [112, 113]. Nevertheless, the identified topics are usually too general. For example, in CrisisLex there is a predefined category of *Infrastructure and Utilities* with words such as damage and road closures, but it provides little information about specific issues. On the other hand, unsupervised methods such as spatiotemporal analyses on tweets have been studied a lot for situational awareness during events or natural disasters [107, 159]. However, they require interactive efforts to determine whether the clusters of topics are relevant.

This study uses a semi-automatic framework to extract disaster-related topics out of the communication contents from citizens and local governments and analyzes the communication patterns at different spatiotemporal scales. First, the framework includes a mechanism that filters out irrelevant information based on the quality of topics generated by the topic model. Second, a topic model is proposed that can explore the topics discussed in a data-driven manner instead of using pre-defined taxonomies. Each tweet is labeled with a topic extracted from the tweets themselves. The geotag tweets with relevant topics are valuable communication footprints for

us to explore the topics discussed by citizens and local governments. Finally, these digital footprints are analyzed from multiple perspectives, such as under the context of the severity of the disaster, or whom to mention (mention other Twitter accounts by using @ in a tweet, simplified as @ behaviors in the following text).

The proposed framework has several contributions to the state-of-the-art: 1) digital communication footprints are identified in an unsupervised manner by the topic model. This type of learning requires minimal human efforts for labeling, and can reveal more insights into the specific types of issues regarding different disasters, instead of commonplace topics from predefined crisis taxonomies; 2) comparative analyses of communication contents between citizens and local governments are conducted at different spatiotemporal scales to identify the potential communication gaps. Specifically, burst analyses on digital communication footprints are utilized to identify the emerging issues specifically relevant to the disasters; topics and themes on the roads are analyzed to show the priorities of users on different types of roads; the @ behaviors are studied to identify whether users have a clear perception of whom to communicate regarding certain issues. The findings can potentially help the local governments to identify citizens needs and make decisions on how to respond under certain conditions during natural disasters.

5.2 The Proposed Method

The framework has four main components (Figure 5.1): (1) data preparation, which describes what kind of tweets to collect and how to clean; (2) tweet selection,

which presents a method on using hashtag tweets as the gold standard to compute the similarity score of geotag tweets and selecting the top similar ones as relevant tweets for topic extraction; (3) topic identification, which extracts and annotates topics communicated online; (4) digital communication footprint analyses, which characterize the distribution of topics at different spatial and temporal scales, and explore communication behaviors under topics.

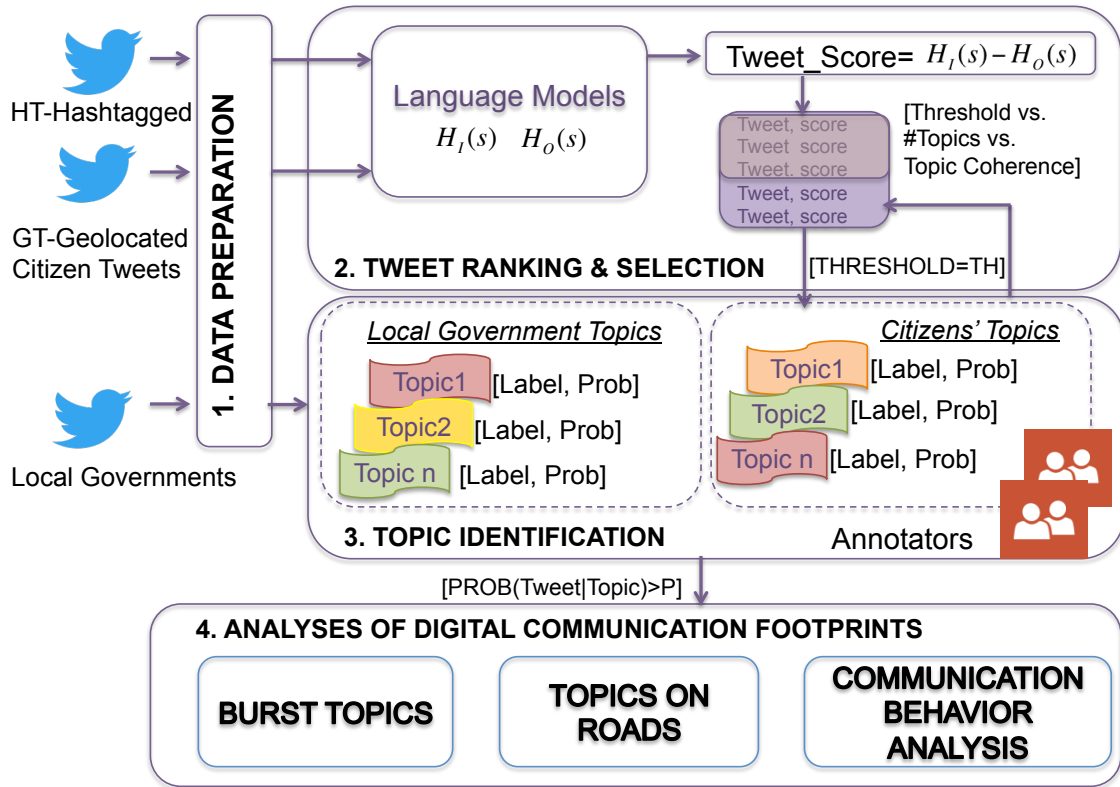


Figure 5.1: Overview of the proposed framework.

The case study in this work is about snowstorms, which usually involve with the closure of schools and government offices, food shortage in preparation, and traffic accidents due to slippery roads. Although cities are usually well prepared, these problems still exist and heavily disrupt citizens. There is a critical need for

a communication channel between citizens and local governments, where citizens report the problems and help to spread the information from local governments, and local governments circulate the preparation and response information to as many citizens in need. Twitter is a widely used platform for such communication. With a better understanding of citizens tweeting content and behaviors, local governments can make better strategies to spread helpful information to citizens in need. The data are collected from snowstorms in Maryland areas from November 2014 to April 2016. There are 18 snowstorms, including some harsh ones with snow depth of more than 10 inches. Snow depth measures the depth of snow on the ground, which is a combination of snowfall and the snow that was already on the ground. Here defines a snowstorm period as one or more consecutive days when the snow depth is at least 2 inches.

5.2.1 Data Preparation

The framework incorporates tweets collected by two methods: hashtag tweets that are related to disaster events, and streamed geotag tweets in the area where the disaster happens. The research question focuses on the local citizens communications to bridge a connection between specific disaster issues and the local governments work. Therefore, the study focuses on geotag tweets with relevant topics, so that problems in local scale can be identified.

The framework analyzes tweets from both citizens and local governments. Local government tweets are collected by scraping information from the web page. The

Twitter handlers representing the local government agencies from the 24 counties in the state of Maryland are collected. In total, there are 83 official accounts and 18,688 tweets in the set. Local government tweets are generally not geotag, but it can be analyzed at the county level based on the account handler.

Citizen tweets are collected by Twitter streaming API using hashtags (denoted as HT) and geo-location (denoted as GT). The snowstorm-related tweets are collected with a list of 12 hashtags that were used during the 2014-2016 snowstorms in Maryland (MD). Instead of generic hashtags such as #blizzard, the chosen hashtags are closely related to the region studied, such as #mdblizzard, #mdsnow. The final dataset contains 96,423 tweets. Geotag tweets were collected as those fall inside the states boundary defined by the state shapefile. Tweets that are not in the time range of snowstorms are eliminated. A snowstorm period is defined by the days when snow depth is at least 2 inches, and two days before and after those days, since there is usually prediction about snowstorms and information about after-effects generated after snowstorm days.

Only individual citizen tweets are taken into consideration in this study. Therefore, accounts that are potentially representing organization or spammers are removed. The final cleaned dataset has 169,987 tweets after these filters.

Before tweet ranking and topic modeling, several preprocessing steps are employed to remove noisy tweets: firstly, stop words, emoticons, and punctuations in tweets are removed. Then I replace specific numbers with the word *NUMBER* and different road names with *ROADNAME* since there are different types of numbers and road names that are frequent in tweets, expressing similar meaning but hard to

model if they were presented as the original format. After that tweets with less than 4 words left are removed since it is usually hard to identify the content of tweets if they are too short. There are 163,019 tweets after the preprocessing. By using the geotag tweets from citizens, it is acknowledged that only the communications from users who have enabled the geo-location are captured. However, only in this way it is possible to extract meaningful local information and analyze tweet contents in a fine-granulated spatial scale.

5.2.2 Tweet Ranking and Selection

Geotag tweets are collected during the snowstorm days; however, there is a significant proportion not talking about the snowstorms. In this part (See Tweet Ranking & Selection in Figure 5.1), I will introduce: (1) a method of ranking geotag tweets by their similarity to the hashtag tweets, which are assumed to be the gold standard as tweets relevant to snowstorms, and (2) a method to choose a threshold that defines the final subset of geotag tweets potentially talking about snowstorms. This selection is exclusively for citizen tweets since the local governments tweets during a disaster typically focus on disaster-related topics [105].

In (1) I use a method called ranking by cross-entropy [160]. It firstly takes all hashtagged tweets (HT) for computing a language model that characterizes the frequency of the n-grams in the text. The output language model is denoted as the *in-domain* model. In contrast to HT tweets, all the geotag tweets (GT) are considered as *general-domain* tweets as they contain some tweets irrelevant to the

snowstorm topics. The same process is applied to the GT to generate the second language model named the *general-domain* model. Then a difference score between the perplexity of the same tweet in the *in-domain* and *general-domain* models are computed. A higher score indicates that a tweet is highly similar to the *in-domain* language but highly dissimilar to the *general-domain* language.

The GT tweets are then ranked based on the difference score. However, how many tweets should be chosen for topic analysis is still a question. Since the final goal is to extract topics related to snowstorms, the threshold is evaluated by measuring the quality of topics. It is assumed that if the selected geotag tweets are snowstorm-related, together with the HT, they generate coherent topics. While if there are many noises, i.e., tweets irrelevant to snowstorms, the extracted topics could be less coherent, since the noisy topics usually are diverse. Therefore, here use the point-wise mutual information score (PMI) to measure the quality and interpretability of the topics extracted by Single-Topic Latent Dirichlet Allocation (ST-LDA). PMI has been shown as the best way to model topic coherence and interpretability [161].

After this two-stage process, the threshold is identified as 20%, and 25 topics are extracted as the combination gives the best PMI score (Figure 5.2). To show that using the rank by cross-entropy difference is better than using a keyword-based approach to select snowstorm-related tweets from all GT, I also compute the PMI values for the keyword-based approach. For the comparative experiment, the keyword list is obtained from two sources: (1) a list of words created by 5 local citizens who are asked to provide as many words as they could for searching snowstorm-related information on Twitter (filtering out repeated words); (2) a commonly used

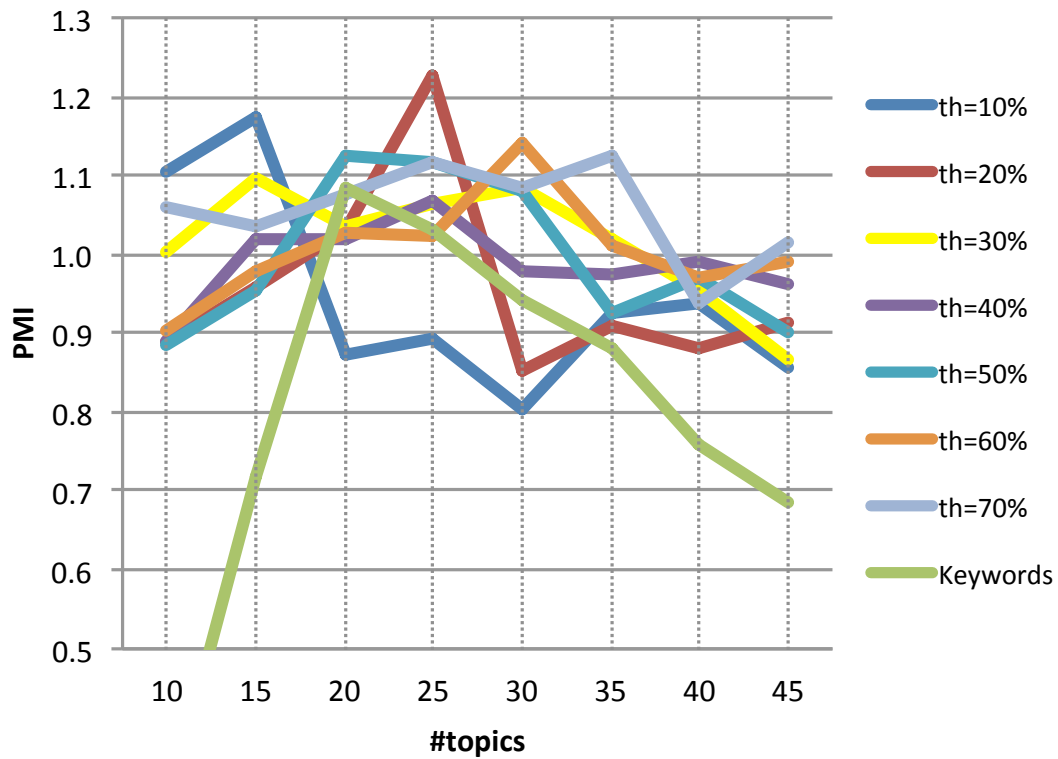


Figure 5.2: PMI values for various thresholds (th) and number of topics (K) and for the keywords approach.

list of 380 disaster-related words [108]. The two sources are combined to create a list of 419 keywords, which are specifically related to the snowstorm events and generally about disasters. I select a subset of GT tweets that contain these keywords, put them together with HT tweets, and run the topic model to obtain the PMI values. The PMI values (grey line in Figure 5.2) are worse than most of the thresholds explored.

5.2.3 Topic Identification

For topic extraction, I apply the Single-Topic Latent Dirichlet Allocation (ST-LDA) model, which is a variant of the topic modeling technique LDA [59]. Different from conventional LDA, which assumes a document with the probabilistic association of multiple topics, ST-LDA considers each tweet as a document and assigns each tweet with only one topic, with a membership probability. Tweets with a membership probability of less than 0.75 are filtered out since they do not have clear topics.

I run two ST-LDA models on citizens tweets and local governments tweets respectively, because the number of local governments tweets is much smaller, and the language of official tweets is usually more formal. If tweets from these two sources were trained together, the topics could be hindered by the citizens topics [162]. To choose the best number of topics for local governments tweets, I explore with different numbers of topics K and use the PMI to evaluate the coherence and interpretability of the topics. The best PMI value is achieved when K is 45. The derived topics are represented by the probabilistic association of words, which is hard to interpret. Four annotators are recruited to help to summarize the content of the topics and assign a label to each topic. Each annotator was given the twenty top-ranking words in each topic, along with a sample of original tweets without pre-processing and the membership probability in that topic. Based on this information, they are required to give a descriptive label for each of the 25 citizen topics and 45 local government topics. The annotators were not offered any guidance about what

labels to assign or any background knowledge about the task. Since there was no communication among annotators, the wording they used could be different, such as *complaints offices delays* versus *office schedule change*. To get consistent labels, a protocol is designed to ask annotators to meet and discuss their labels, and evaluate whether they are referring to the same topics. For these topics, they achieved a final agreement on the names of the label. After that, the inter-annotator agreement using Fleiss Kappa [163] is 0.78. For topics without full agreement, the label achieves agreement over more people is used. Some topics are highly related and thus are grouped into themes, for example, *traffic accidents* and *driving conditions* are categorized into the theme of *traffic*. Finally, there are seven themes identified.

| Themes | Topics | Description | Top Ranked Tweets |
|--------------------------------|-----------------------------|---|--|
| Economic and Social Factors | Economic and Social Factors | Thoughts and opinions around the impact of snowstorms including the impacts on work, money and food supplies in stores etc. | Sears has no love for their employees making me work in this bullshit weather |
| Government Offices and Schools | * Delays and Closures | Information about delays and closures of school or offices | @FCPSMarylandd: Frederick County Public Schools will open 2 hours late Friday Feb 20th, 2015 due to weather. @Ary_Nikporaa |
| | Complaints | Complains about schedule changes of offices or schools, or suggestions that schools should be closed given the harsh weather conditions | Mcps is the type of county that would still give school even with this weather on Tuesday |
| Public Transportation | * Delays and Closures | Schedule changes or delays of public transportation such as flight, train, metro and buses | 20 hr travel day back to Europe starting w a 2 hr delay on 1st flight ... excellent ... NOT ... LOL. |
| | Private Options | Uber, Lyft, and price surges during snowstorms | How did I get an Uber_Maryland today without Surge pricing? [...] Thank you Uber |
| Traffic | * Accidents | Report of car crashes and road accidents | I'm at Accident, MD in Accident, MD [link] |
| | * Driving Conditions | Report of slippery road or snow on road | Roads lookin ridiculously unsafe @MCPS |
| Response | * Snow Removal | Tools, or progress of snow plowing | Thank you @MCPS for not plowing your parking lots |
| Weather Info | * Weather Info | Information/pictures about weather conditions | Old Village and Mechanicsville Road (picture was taken around 8:30 a.m.) this morning |

Table 5.1: Themes and topics from citizen tweets, * indicates common topics for both citizens and local governments.

Table 5.1 and Table 5.2 show the themes and topics in citizens and local governments respectively. Although models are trained separately on citizen tweets and local government tweets, there are common topics, which are labeled the same,

| Themes | Topics | Description | Top Ranked Tweets |
|--------------------------------|----------------------|--|---|
| Government Offices and Schools | *Delays and Closures | Notice of school or offices delays and closures | UPDATE: Anne Arundel County Government Offices Will Be closed Today |
| Public Transportation | *Delays and Closures | Schedule changes or delays of public transportation such as flight, train, metro and buses | RTA bus service remains suspended & will not operate 1/25 [...] for service updates visit |
| Traffic | * Accidents | Report of car crashes and road accidents | Serious Accident: Road shut down on Veterans Hwy between Brightview and Eastwest. Pleas avoid area. |
| | * Driving Conditions | Report of slippery road, road close or open | Update: #MdTraffic: Whites Cove Rd @ Ft. Smallwood Rd is reopened. Drive "Safely"! |
| Preparedness | Preparedness | Safety tips or advisory information of how to get prepared for the snow storm | Protect your pipes against freezing by letting a thin stream of water flow through faucet |
| Response | * Snow Removal | Tools, locations and times and progress of snow plowing | 98% of County roads plowed at least once. Constant work continues to the 2,600 miles [...] |
| | Emergency Services | Offering ways citizens can ask for emergency help, the process of coping with emergency calls. | MCFRS Emergency Communications on 'CONDITION RED' due to heavy call volume [...] |
| Weather Info | * Weather Info | Weather conditions such as snow depth, temperature | Snow totals for AACO are up. Now in the 8-10 inch range |

Table 5.2: Themes and topics from local governments tweets, * indicates common topics for both citizens and local governments.

and indicated in the tables with a * in front. The theme of *economic and social factors* is exclusive to citizens. It is a mixed theme about different impacts of the snowstorm on citizens life, such as companies forcing employees to work under harsh weather conditions; food supply shortage in stores; working from home in bad weather. The theme of preparedness is uniquely formed in local government tweets.

Both local governments and citizens use Twitter as a platform to notify schedule changes of schools or offices in the theme *government offices and schools*. Meanwhile, citizens have many complaints about schedule changes or suggest that schools should be closed given the harsh weather conditions. The words in the *delays and closure* topic are mainly about times and places. While tweets labeled with the complaints topic usually include trending words such as closefcps and usually describe weather condition and delay or closure at the same time.

In *public transportation*, the topic of *private option* is exclusive to citizens, which talks about surging prices of share rides such as Uber and Lyft. It raises a question as what local governments can help to solve travel problems of citizens who usually rely on public transportation.

In the theme of *response*, the topic of *emergency services* is only shown in local government tweets, which is mostly about offering ways in which citizens can ask for emergency help and the process of coping with emergency calls.

Although both citizens and local governments share the theme of *weather info*, the top-ranked tweets show that citizens mostly describe weather conditions or share photos about the weather, while local governments usually give accurate information about snow depth, temperature.

5.3 Results

5.3.1 Spatiotemporal Burst

The purpose of spatiotemporal analyses is to understand the emerging issues during snowstorms, i.e., the topics that attracted the attention of citizens of a closely connected area in a short time. Here I apply the Space-Time Permutation Scan (STPS) Statistics on citizens communication footprints to identify the significant hot-spots [164]. The method is originally designed for detecting disease outbreaks. It uses multiple overlapping cylinders as candidate windows, where the base represents a geographic area, and the height represents the number of days. The geographic area is searched by iterating over all geographic units with the circle radius from zero to some predefined maximum value. For each location and size of the window, the numbers of observed and expected cases are counted. Among these, the most unusual excess of observed cases is noted.

In this study, every geotag tweet with the topic related to snowstorms is con-

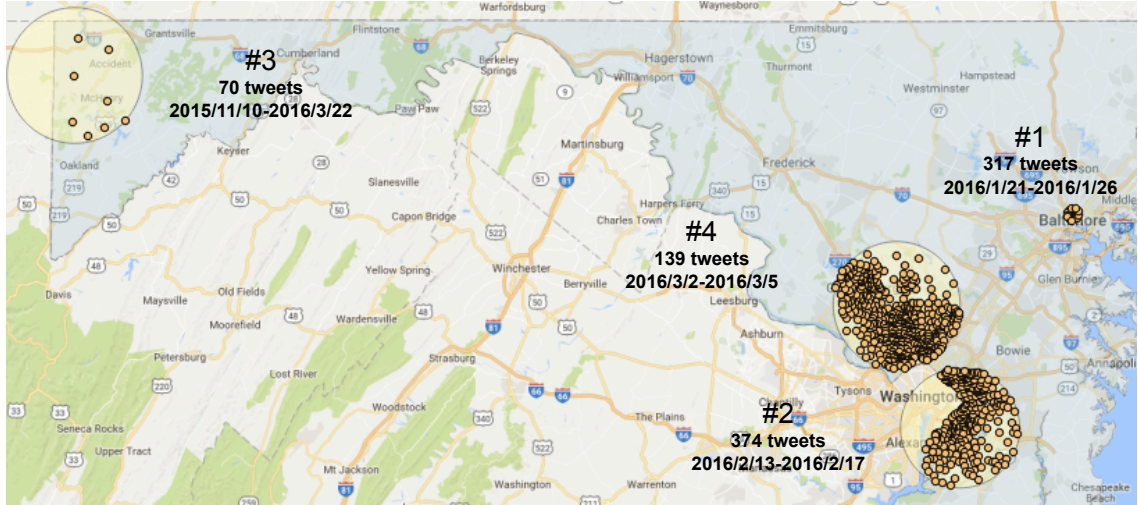


Figure 5.3: Spatiotemporal clusters detected by STPS.

sidered as a case. To keep the fine-granulated spatial information, I choose the Census Block Groups as basic geographic units for cluster detection, i.e., each tweet is labeled with the Census Block Group it falls in as location.

Figure 5.3 shows the detected four hot-spots in Maryland (covered by blue) located in: 1) the center of Baltimore City; 2) Prince George County, Northwest Washington D.C. (DC); 3) Garrett County and 4) Montgomery County, North of DC. For each cluster detected, the percentage of themes are computed.

Local government tweets are usually not with geotags. However, the location can be identified by the account information. For each cluster, I retrieve the official tweets from the same county within the same time range to compute the themes of communications. I also compute the themes for all the other tweets that are not clustered in hot-spots as a baseline.

Figure 5.4 shows the identified themes for citizens and local governments. In the baseline group, the theme *economic and social factor* takes a substantial propor-

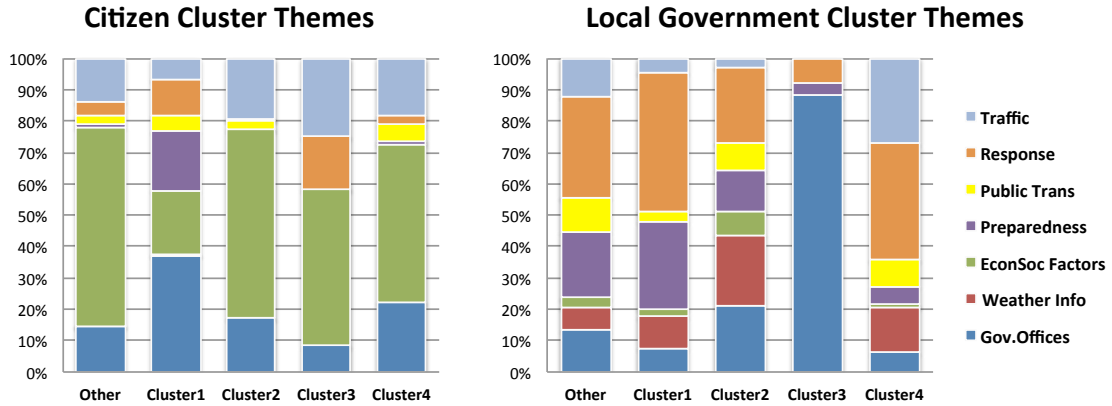


Figure 5.4: Themes of tweet clusters by citizens and tweets by local governments in the same spatiotemporal scale.

tion in citizen tweets, while local government tweets are more about *preparedness and response*. There are more tweets about *traffic* (13.69%) than *public transportation* (2.42%) from citizens, which presumably because more citizens suffered from the impact of snowstorms on traffic than public transportation. While generally, local governments published more information about *public transportation*.

By comparing the themes in hot-spots and the baseline, the prominent issues can explain why the hot-spots are formed. Of the four clusters, Cluster 1 is in the Baltimore City with urban form. It shows a major theme, i.e., *government offices and schools* (36.93%), of which, 34.66% is about delay or closure information, while 2.27% are complaining tweets. These numbers show citizens care about this issue and have the motivation to spread the information, however, only 7.66% of the tweets from the local government are about schedule changes of government offices and schools. It shows insufficient communication directly from local governments to citizens, assuming that if the issue is well communicated, the hot-spot cluster

would not be formed. Meanwhile, the local government has a heavy focus on snow removal information, which takes a proportion of 34.99%. The difference seems to reflect that local government has a mismatch for citizens needs.

Cluster 2 in Prince George County and Cluster 4 in Montgomery County have similar theme proportion, which could due to the similar urban landscape and the developed public transportation system. They both located in the surrounding DC area, where the beltway I-495 and highway I-95 go through. These two clusters have the themes of *response* and *public transportation* similar to the baseline. Most noteworthy, the theme *traffic* is prominent in these two clusters. It accounts for 19.29% in Cluster 2 and 18.42% in Cluster 4, which is higher than 13.69% in the baseline. Of the traffic tweets in Cluster 2, 55.52% are about accidents. In Cluster 4, the proportion is 78.56%. They both have a much higher proportion of accident topic in *traffic* theme compared to the baseline, which is 15.49%. It shows that when there is a heavy snowstorm, highways in these two areas have a potential higher risk of car accidents, which should be paid attention.

Cluster 3 in Garret County mainly locates in suburban and rural areas. The theme pattern shows significant differences from that of Cluster 2 and 4. In this cluster, citizens do not discuss *weather info*, *preparedness* or *public transportation*. On the contrary, *traffic* and *response* tweets take a much more significant proportion than the baseline. Unlike Cluster 2 and 4, where most *traffic* tweets are about accidents, the major traffic tweets in this cluster are about driving conditions. The local government tweets only cover the themes of *government offices and schools* (88.24%), *preparedness* (3.92%) and *response* (7.84%). The large proportion of traf-

fic theme in citizen tweets and the lack of traffic information from local governments reflect the differences in communications. It might indicate a need for better road conditions during snowstorms from the citizens side, which local government could pay more attention to.

In summary, using spatiotemporal cluster analysis, I detected four clusters that indicate different issues during snowstorms. The cluster in Baltimore City shows a mismatch between the citizens information needs and the communication contents in local government accounts. The clusters in Montgomery County and Prince George County show the issues urban citizens usually concern about, which have a focus on the information on the *traffic, public transportation, and government offices and schools*. Comparatively, the cluster in Garret County reflects different issues to be prioritized for suburban and rural areas, which is *traffic and response*. Local governments may take actions accordingly during snowstorms.

5.3.2 Themes and Topics in Roads

Figure 5.5 shows the digital communication footprints for local governments and citizens on roads. The objective is to understand the topic distribution for a given spatial scale, and to compare across local governments and citizens. To carry out the analysis, the framework assigns each geotagged tweet in *CT* to the county whose geographical boundaries contain the geolocation of the tweet; and to the road whose boundaries, with a $2m$ buffer, include the location of the tweet. TIGER/Line shapefiles are used to extract the exact boundary information [165] for

the 24 counties and for five types of roads: interstates, freeways, principal arterials, minor arterials and collector streets. Local government tweets in *LT* are not geotag and as a result are assigned to the county of the agency represented by the twitter handle; or to roads (without a specific type) if the twitter account is from a local transportation agency.

The first two bars on the left, represent the average footprint across all types of roads for local government (*LT*) and citizen (*CT*) communications. We observe that, similarly to the county communication footprints, local government communications focus on *preparedness* and *response*, possibly offering pre-snowstorm advice and updates about snow removal strategies. However, we also observe that the official discourse on roads puts a lot of weight on communications about *traffic* and *public transportation*, with tweets spreading details about accidents or driving conditions; and offering information about delays and closures of public transportation services. On the other hand, citizen communications on roads still focus on *economic and social factors* (35%), and *government offices and schools* (14%). However, citizen traffic-related communications have almost doubled with respect to their county communications, from 18% to 35%. Looking in depth into the citizens' communication footprints for each type of road (Figure 5.5, five bars to the right), we observe that the distribution of topics changes from larger roads (interstates, *INT.* or freeways *FRE.*) to smaller roads (principal and minor arterials *PRI.,MIN.* or collectors *COL.*). Interestingly, while citizens mostly focus on sharing accident information and driving conditions with others when they are on large interstates or freeways (65% average communications on the traffic theme), the discourse changes towards

economic and social factors (48%) when we move to smaller arterial and collector roads. This reflects that citizens' communications during snowstorms change depending on the physical spaces they occupy. Driving in large roads during snowstorms encourages a utilitarian approach with drivers sharing useful information (Twitter as a news platform) while smaller roads bring them back to opinion and thought sharing (Twitter as a social network).

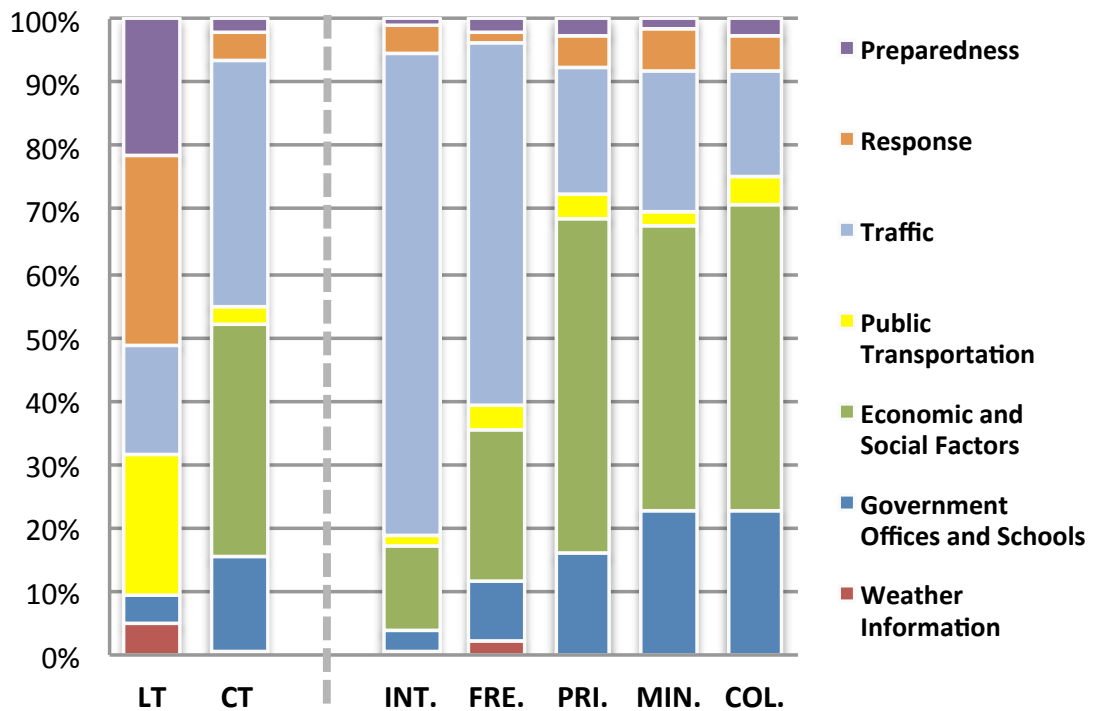


Figure 5.5: Distribution of themes and topics on roads.

5.3.3 Interactions between Citizens and Local Government Accounts

I use the digital traces labeled with themes to analyze the interaction behaviors between citizens and local governments regarding different issues. In Twitter, citizens use the user mention function (mentioning other Twitter users in the text

by using the symbol @ with a Twitter account, simplified as @ function) to communicate with certain accounts directly. Different from the text that shows in the news feed of the followers page, the @ function is used as a direct communication channel between citizens and local governments.

| Themes | Number of citizen tweets | Number of tweets have @ (percentage) | Total number of @ | Times of government accounts being @ |
|-----------------------------|--------------------------|--------------------------------------|-------------------|--------------------------------------|
| Economic and Social Factors | 14001 | 4546(32.47%) | 6427 | 37(0.58%) |
| Gov. Offices and Schools | 3283 | 980(29.85%) | 1335 | 191(14.31%) |
| Public Transportation | 545 | 221(40.55%) | 275 | 9(3.27%) |
| Traffic | 3049 | 796(26.11%) | 1034 | 58(5.61%) |
| Preparedness | 347 | 94(27.09%) | 124 | 6(4.84%) |
| Response | 1031 | 348(33.75%) | 465 | 36(7.74%) |
| Weather Info | 18 | 2(11.11%) | 2 | 0(0%) |

Table 5.3: The @ behavior of citizens under various topics

Table 5.3 shows the number of citizen tweets under each theme and the analysis of @ behaviors. On average, there are 31.37% tweets with @ behavior, while only a small proportion (3.49%) refers to local government accounts. Among the topics, citizens mention local governments more for the theme of *government offices and schools*. Government accounts are being mentioned 191 times, which takes 14.31% of all. On the other hand, government accounts are mentioned the least in the theme of *economic and social factors* (0.58%) and *public transportation* (3.27%), which may be perceived as less related to local governments by citizens.

The top mentioned accounts in *economic and social factors* include @BarackObama, @EACgov, @BrubeEnberg, @Bexofeasttex, and @babyspittle, which are accounts of politicians, political organizations and economists. The tweets mentioning these accounts are mostly about funding allocation for public services and government responsibilities in disaster relief.

The top mentioned accounts in *government offices and schools* are @MCPS,

@FCPSMaryland, @pgcps, @AACountySchools, and @fox5newsdc, which are mostly education systems and news media. It shows that citizens have a clearer perception of whom to ask for help or to report issues regarding the theme of *government offices and schools*, comparing to other issues.

For *public transportation*, citizens mostly mention accounts of departments that are in charge of schedule management during snowstorm days such as @Amtrak, @wmata, or specific accounts that report public transportation information or citizen complaints about the metro, e.g., @unsuckdcmetro, @unsuckdcmetr. Comparing to accounts handled by local governments, there are information hubs that are more popular or more active in communications with citizens.

The frequently mentioned accounts in the topic of *traffic* are @EmpireFOX, @MCPS, @MCPSsnow, @vbalradio, and @WTOPtraffic. Although there are government accounts, such as Sheriffs in counties, and Departments of Transportation working on traffic issues during disasters, citizens appear not to be aware of those accounts or not to be motivated to communicate with them. Comparatively, citizens prefer to spread traffic information through news media outlets or radio stations. For *preparedness* and *response* related tweets, people also tend to mention news outlets such as @fox5newsdc, @cbsbaltimore and @capitalweather from the Washington Post.

The analysis shows that citizens usually do not have a clear idea about the government accounts they can communicate with through Twitter. Although the @ behaviors are common over almost all themes, the frequency local government accounts being @ is rare. The exception is for the theme of *government offices and*

schools, in which the @ behavior to local government accounts is much higher. From the top mentioned accounts, we see that citizens prefer media accounts and super users that act as information hubs in the Twitter platform for communication.

5.4 Limitations

The limitation of this study lies in the following perspectives: First, the geotag tweets are a biased representation of all citizen tweets. The geotag tweets for topic analyses can be a small proportion of all tweets posted by local citizens. Considering that the location service users are biased towards younger users, users of higher income, and users in urbanized areas [166], the result reflects more concerns from this group of people.

Second, the theme composition is computed based on the assumption that digital traces are equally weighted, which to some extent reflects the focus of on-line communication by citizens and local governments. However, to evaluate the efficiency of local government communication, it would be more important to understand the impact, i.e., the number of citizens who received the necessary information in time. The impact of each tweet can be different depending on the popularity of the account that published it and the way people spread it online. A piece of information might achieve large audience coverage if citizens participated in the spreading of the information.

Chapter 6: Study 3: Identification and Characterization of Internal Migrants with Cell Phone Data

6.1 Introduction

Internal migration refers to the migration of individuals from one region to another within the same geopolitical entity, typically within the same country [167, 168]. There has been an increase in the volume, types, and complexity of human internal migration in recent years [169]. Considerable attention has been given to the study of the determinants and consequences of internal migration using macro-level and micro-level analyses. Macro-level studies are carried out using a combination of various survey and census datasets, including origin-destination internal flows as well as demographic and socio-economic data, to assess the role that specific variables might play as both determinants or consequences of internal migration movements [120]. However, such methods are unable to reveal the nuanced causes and consequences of individuals' migration behaviors, which are important information for the development of policies that aim at the intervention of individuals' adaptation process. Micro-level analyses reveal the physical and psychological status of migrants after their migration by using interviews and surveys. The data

acquisition process is usually expensive, making it hard to scale to a large population. On the other hand, cell phone data enable us to look into the behavior traits of individual migrants of a large scale [170]. Due to the geographical scale, the data also enable the examination of internal migrants with other contextual information, for example, to compare migrants in urban or rural settings. Since cell phone metadata contains the georeferenced locations visited by an individual and her contacts, it allows to model individual behaviors both in terms of spatial dynamics as well as social networks. In this study, I present a framework that uses information extracted from cell phone metadata to reveal internal migration behaviors that could guide or complement the research agenda of micro-level migration researchers working to understand the physical, social and psychological decision processes behind migration experiences.

The proposed framework allows to reveal internal migration behaviors with a focus on immediate post-migration behaviors and the role of pre-migration activities from two perspectives: spatial dynamics and social ties. The main objective is to carry out large-scale analyses of internal migration trends so as to reveal individual migrant behaviors that would benefit from further qualitative studies through personal interviews or individual surveys. Ultimately, we expect the analyses to inform migration researchers of pre- and post-migration behaviors that would benefit from further qualitative analysis.

The proposed framework consists of three parts. First, the framework uses features extracted from the cell phone metadata to identify potential migrants in the dataset. I present a method to identify internal migrants and evaluate its ac-

curacy using real census migration data. Second, the framework uses the identified migrants to characterize immediate post-migration behaviors *i.e.*, analyzing the spatial dynamics and social networks of migrants post-migration, and compare these against behaviors from locals that have not undergone any migration process. Third, I analyze the role that pre-migration spatial dynamics and social networks might play in the same post-migration behaviors shown by internal migrants. I evaluate the proposed framework to study internal migration behaviors in Mexico, using a dataset with eight months of anonymized cell phone metadata from over 48 million subscribers; and I show how the findings could complement future qualitative studies in Mexican internal migration.

6.2 The Proposed Method

There are limitations of survey data: the survey samples are usually limited and sometimes biased, and the results usually depend on respondents' intentions rather than the actual deeds [171]. While, the cell phone data can be used to study internal migrants on a large scale, thus overcome the potential limitation of sample size if using survey methods. The CDRs are automatically generated, which means the study is easily scalable. Meanwhile, the CDRs reflect multiple dimensions of users behaviors including spatial dynamics and social networks thus enables exploration of behaviors from multiple perspectives.

I present a framework that uses cell phone metadata to analyze internal migrant behavior in the immediate post-migration period quantitatively and to

evaluate the role that pre-migration behaviors might play on the immediate post-migration activities of internal migrants.

The analyses focus on two types of behavioral features: spatial dynamics and social ties. Spatial dynamics refer to the mobility patterns that individuals have. These variables can reveal insights for the spatial dispersion or spatial diversity of migrants, throwing light into the impact that migrations might have in how individuals explore their new physical environment. On the other hand, social features provide information about the social ties of migrants, before and after migration, which might help to analyze the role that social relationships might play in the immediate adaptation of migrants to their new environment. To analyze the spatial dynamics and social networks of internal migrants, we first need to identify the migrants themselves in a cell phone dataset. I present a method that uses cell phone metadata to identify long-term changes in home locations and evaluate its accuracy against real, census-based, internal migration data. The framework will contribute with quantitative methods to answer the following two research questions (see Figure 6.1 for details):

RQ1. Characterization of immediate post-migration behaviors. I present a method to analyze spatial dynamics and social networks in the immediate post-migration period and to compare these against local individuals to assess the behavioral differences potentially due to the migration. This analysis will reveal insights that might be helpful for future qualitative studies focused on understanding how internal migrants adapt to their new settings, and on the difficulties that they might find in that adaptation process.

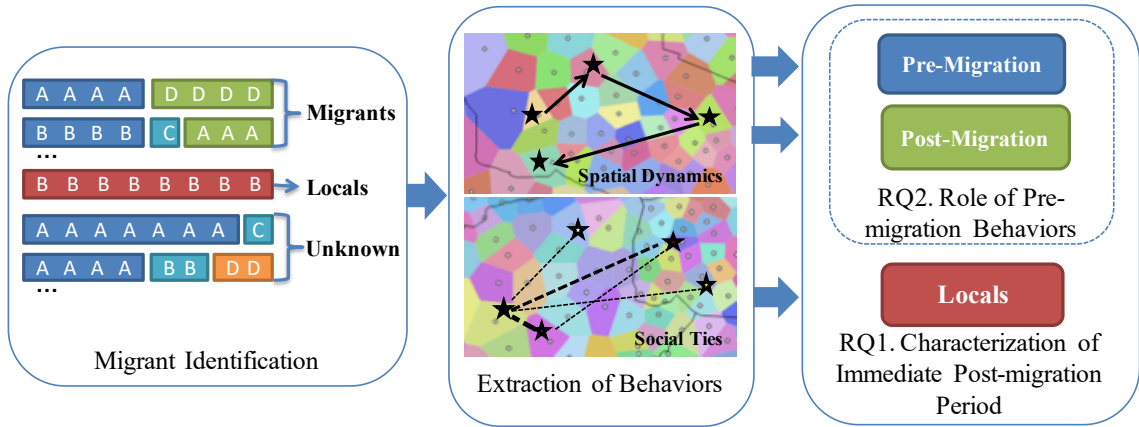


Figure 6.1: Overview of the proposed framework and the two research questions it answers: Identification of migrants, (RQ1) Analysis of the behavioral consequences of internal migrants and (RQ2) Analysis of the potential causes of the post-migration behaviors. RQ1 and RQ2 use two types of behavioral traits extracted from the data: spatial dynamics and social ties

RQ2. Analysis of the role that pre-migration behaviors have on post-migration activity. I describe a multivariate regression method to examine and quantify the role that the pre-migration spatial dynamics and social networks have in the post-migration behaviors that migrants show immediately after they arrive at their new locations. The model can potentially be used as a tool to foresee the types of spatial and social behaviors that internal migrants might have, given a specific type of pre-migration population. Further qualitative research in this direction, could translate into the development of policies to better assist and ease the adaptation of migrants to their new surroundings [172].

I explore the applicability of the proposed framework and methods using cell phone metadata for the country of Mexico, and present behavioral findings that

illuminate the spatial dynamics and social relationships that internal migrants have.

6.3 Data Description

I use two main data sources, cell phone traces to characterize internal migrants, and census information to evaluate the detection of migrants.

6.3.1 Cell Phone Data

To characterize migrant spatial dynamics and social ties, I use the information collected by a cell phone network. Cell phone networks are built using a set of base-transceiver stations (BTS) that are responsible for communicating cell phone devices within the network. Each BTS or cellular tower is identified by the latitude and longitude of its geographical location. The coverage area of each BTS is called a cell. The geographical area covered by a cell mainly depends on population density, and typically ranges from less than 1 km², in dense urban areas, to more than 4 km², in rural areas. For simplicity, it is common in the literature to assume that the cell of each BTS is a two-dimensional non-overlapping polygon, which is typically approximated using Voronoi diagrams [71, 73].

CDRs are generated by telecommunication companies for billing purposes. CDRs can be used to model spatial dynamics and social ties. Spatial dynamics are characterized by features that use the geographical position of the cellular towers used for placing calls, while social ties can be characterized using the people that individual talks to. Sections 6.5.1 and 6.5.2 will describe in depth the specific

features.

Here, I use an eight-month anonymized CDR dataset for the country of Mexico (October 2009 to May 2010). To preserve privacy, original records are encrypted, and all the information presented in the study is aggregated. From all the information contained in a CDR, the study considers the encrypted originating number, the encrypted destination number, the time and date of the interaction, and the BTS that the cell phone was connected to when the call was placed.

The dataset contains 7 billion records from 39K cellular towers that cover the whole country. I eliminate from the dataset all cell phones (and their corresponding CDR data) whose activity can be assumed to correspond to a machine and not an individual using the approach in [33]. This approach, which uses average measures of reciprocal cell phone contacts and frequency to eliminate anomalous accounts, was applied over the dataset leaving a final number of 48M unique cell phones in the dataset.

6.3.2 Census Data

The census data to evaluate the identification algorithm is obtained from the Mexican Statistical Institute (INEGI), a migration matrix at the municipality level. The migration matrix is based on the ENADID 2010 survey (National Survey of Demographic Dynamics) [173]. It records the number of people migrating from one municipality to another from 2005 to 2010 across the whole country.

6.4 Identification and Validation of Internal Migrants

In this section, I present a method to identify internal migrants in a CDR dataset automatically; and propose two approaches to validate that the migrants and migration flow identified can be used as representatives of the country migration flows at large. I carry out the validation by comparing the internal migrants, and their migration flows against actual census data and report their similarity.

6.4.1 Identification of Internal Migrants

Since we only have an eight-month observation window with CDR data, I define as internal migrants individuals that have a consistent home location for at least three months and then move to another place, where they also stay for at least three months. A similar method was described by Blumenstock *et al.* in [134]. With this definition, the internal migrants can be either long-term or short-term (circular) migrants depending on whether they go back or not to their original location sometime after our data collection period finishes [174]. A limitation of our work is that I did not differentiate between these two types of internal migration. However, since I compare this method against internal migration census data that does not differentiate between the two, the validation is still consistent.

Home location algorithms are contained within a larger group of algorithms used to identify the important or meaningful places of an individual from their mobility information. Although the bulk of state-of-the-art focuses on CDR, the algorithms can be used for any set of non-continuous location traces. In the literature,

important places are typically classified as home, work or other [175]. The main idea behind these algorithms consists in using some criteria to define time slots for home, work, and other activities and then use the mobility information available to identify the location of these important places. The most well-known approaches include Ahas *et al.*, who used an anchor-point model to identify home and work and validated it with the actual geography of the population finding a high level of correlation [176]; the work by Isaacman *et al.* who clustered cellular towers (or active points) to identify home and work [177]; or the work by Frias-Martinez *et al.* who proposed a genetic algorithm to identify the time slot that had to be used to better characterize home and work [178].

Based on state-of-the-art described, I identify the daily pre- and post-migration home location of an internal migrant as the most used BTS tower between the hours of 6 pm and 6 am Monday through Thursday each week, with the assumption that it is highly probable that a person is at home at night and on weekdays. We also explored other home location methods based on the center of gravity of the most used BTS towers as presented in [134, 177]. However, the migration validation that I discuss in this section showed that the best results were obtained for the time-range method.

Finally, I assign home locations at the municipality level because the census data that I use to validate our method measures internal migration at that granularity level. To identify internal migrants and their flows, we need to determine the individuals whose home location was at a given municipality for at least three months, and then changed to another municipality and remained the same for at

least another three months. To achieve that, I compute home location (municipality) as explained, for each on a daily basis. If there is no information to identify a home during a weekday, the last position identified for weekdays was assigned. However, any user with at least one week without a home location assigned is not considered as potential migrant due to a lack of mobility information. Once all daily home locations have been computed, if at least 70% of the municipalities identified as home location are the same, the individual is assigned that home for the month. Finally, I search for individuals whose municipality home location throughout the eight months of data changes only once, with at least three months in each municipality. Following this procedure, I identify a total of 18,580 internal migrants in the dataset.

6.4.2 Validation of Internal Migrants

We conduct two validation experiments to verify the consistency between the migration flows detected with the above method, and the real migration flows.

Validation 1. I use the *real* migration flow matrix computed by the Mexican Statistical Institute (INEGI) which is based on the ENADID 2010 survey (National Survey of Demographic Dynamics) [173]. The survey records the number of people migrating from one municipality to another from 2005 to 2010 across all municipalities in Mexico. I use the internal migrants identified by our method to compute a CDR-based migration flow matrix, and compare it against the official one via Pearson’s correlation analysis between the two matrices. Specifically, I com-

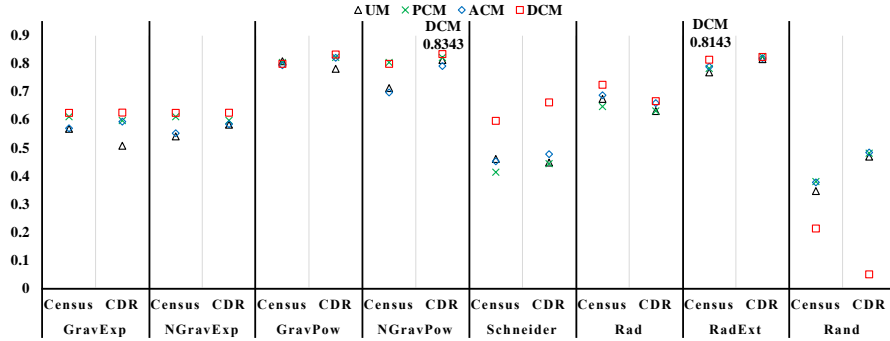


Figure 6.2: Goodness of fit with census-based and CDR-based Migration Flows.

The migration models explored are GravExp, NGravExp, GravPow, NGravPow, Schneider, Rad and RadExt. The constraints are UM, PCM, ACM and DCM.

pute the correlations for: (i) the migration between each pair of municipalities *i.e.*, correlation between each pair (origin,destination) in the flow matrix, (ii) the out-bound migration across municipalities *i.e.*, correlation between each pair (origin,all destinations) in the migration flow matrix and (iii) the inbound migration across municipalities *i.e.*, correlation between each pair (destination, all origins) in the migration flow matrix. The results show a strong correlation between the real and the CDR-based matrices with correlation coefficients of .60, .82 and .74, respectively (with all $p < 2.2e - 16$).

Validation 2. The previous validation compared the official internal migration flows across five years against the CDR-based migration flows obtained with eight months of data from 2009-2010. As such, the correlation coefficients might be affected by the disparity in the data collection periods. To overcome this limitation, I present a second validation that, rather than comparing actual migration behaviors against CDR-based migration behaviors, compares the similarity between

the outcomes of migration models and the two datasets. By comparing similarities between model outcomes and actual migration volumes, we expect to minimize the impact of diverse temporal windows in the analysis. There exist two main families of migration models to explain internal or international migration patterns: Gravity models [179] and Schneider’s Intervening opportunities models [140] (the latter have evolved into Radiation models [180]).

While gravity models assume that the volume of migrations between two locations decreases with distance, intervening opportunities models assume that the migration flow between two locations depends on the number of opportunities that each location has *i.e.*, the decision to migrate is not related to the distance between the two places, but rather to the possibilities of settlement at the destination location. The two types of migration models have been used in the past to explain both migrations and commuting patterns with various levels of success [180–182].

I fit these two types of models using: (i) distances between origin and destination municipalities for the gravity models; and (ii) population of the municipalities [183] as a proxy for intervening opportunities (as is currently done in state-of-the-art for migration models [180]). Once these models are built, I compare its *theoretical* outcome against the census-based migration flows, and the CDR-based migration flows and reports prediction accuracy. I measure similarities between the models and the flows using the common part of commuters feature (CPC) [181] which measures the percentage of correct predictions for the number of people that migrate to other municipalities. It varies from 0, when no agreement is found, to 1, when the two migration flows are identical. Specifically, I fit the following

migration models: Gravity law model with an exponential distance decay function (*GravExp*); Normalized gravity law model with an exponential distance decay function (*NGravExp*); Gravity law model with a power distance decay function (*GravPow*); Normalized gravity law model with a power distance decay function (*NGravPow*), Schneider’s intervening opportunities law model (*Schneider*); Radiation law (*Rad*) and Extended radiation law models (*ExtRad*). Each one of this migration models considers four different constraints: Unconstrained model, or *UM*, that just preserves the total number of trips; Production constrained model, or *PCM*, that assumes that the number of trips produced by a geographical unit is preserved; Attraction constrained model, or *ACM*, that assumes that the number of trips attracted by a unit is preserved and Doubly Constrained Model, or *DCM*, that assumes that both the trips generated and attracted are preserved (see [181, 184] for further details).

Figure 6.2 shows the CPC results for both the census-based migration flows and the CDR-based migration flows for each migration model and each constraint. The extended radiation law model (RadExt) [185] is the one that performs the best with a relatively high goodness-of-fit of 0.8143 for the census-based migration flow; while for the CDR-based migration flow, the normalized gravity law (NGravPow) performs best with a goodness-of-fit of 0.8343. These results indicate that the set of internal migrants identified using the method proposed in this section appears to represent well the real internal migration flows at the country scale. Next two sections will use these internal migrants to analyze immediate post-migration behaviors and the role of pre-migration behaviors

6.5 Results

With the detected migrants, I conduct the following analyses to characterize the behavior consequences of internal migrants and how their pre-migration behaviors relate to their post-migration activities.

6.5.1 RQ1: Behavior Consequences of Internal Migrants

In this section, I analyze the internal migrant behavior in the immediate post-migration period from two perspectives, Such comparison will allow us to identify behavioral differences that could be potentially explained as consequences of the internal migration process, rather than as artifacts of the new physical environment *i.e.*, a migrant might change her commuting patterns post-migration, but this could be due to the fact that one is migrating to a city with a different urban geography. Related work has shown that spatial dynamics and social networks are often shaped by their physical environment [186–188]. Thus, by looking at differences with local behaviors, we expect to discern between changes due to the physical environment or issues that migrants face and that are specific to their migrant community. We identify the local population (locals) in the CDR dataset as individuals who have a home location assigned to a municipality where internal migrants migrate to, and whose home location is the same throughout the length of our dataset (eight months). Using the home location algorithm explained before and these requirements, there are 1,505,868 local residents across all municipalities that I have identified as migrant destinations. To assess that the local population sampled as local can be used as a

representative of the population at large, I compute Pearson’s correlation between the real population numbers per municipality and the population sampled as local for that municipality. The two values are consistent across most municipalities, with a high Pearson’s correlation coefficient of 0.8570, and a Spearman’s coefficient of 0.8295 (with $p < 2.2e - 16$ for both), showing a high-rank correlation as well.

Our final objective is to analyze the immediate post-migration behaviors of migrants and compare these against the behaviors of locals to assess migrant behaviors potentially the consequence of the migration process itself and not of the physical environment they migrate. I carry out this comparison by modeling spatial dynamics and social network features for both migrants (pre- and post-migration) and locals with the CDR data available. Next, I identify the migrant features that show statistically significant different values between the pre- and post-migration periods, reflecting behavioral changes. After that, to discern between behavioral changes due to the new physical environment or to the migration process itself, I compare the post-migration migrant features against the same features computed for locals. Statistical analyses between the two distributions will allow us to identify behavioral differences that could be attributed as consequences of the internal migration movement.

Spatial dynamics We consider the following features to characterize individual spatial dynamics regarding spatial dispersion, diversity, and entropy:

- 1) *Number of Municipalities Visited*. It is computed as the monthly average number of municipalities visited by an individual. This feature is calculated identifying the municipalities that correspond to the cellular towers where a given individual


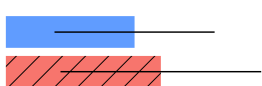
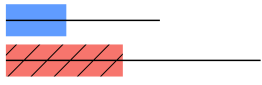

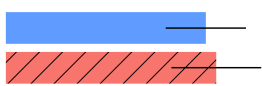
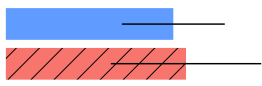
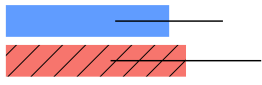
| Features | Mean and SD after Box-Cox  | Welch's t-test | | Cohen's d | |
|--|--|----------------|----------|-----------|-------------|
| | | t | p-value | d | Effect Size |
| #1. Number of Municipalities |  mean=1.66 sd=1.03 mean=2.00 sd=1.29 | 34.3860 | <2.2e-16 | 0.3189 | small |
| #2. Entropy of Municipalities |  mean=0.14 sd=0.22 mean=0.28 sd=0.33 | 54.4800 | <2.2e-16 | 0.5654 | medium |
| #3. Daily Mobility <i>Box-Cox</i> $\lambda=0.1414$ |  mean=1.17 sd=0.22 mean=1.21 sd=0.24 | 16.2850 | <2.2e-16 | 0.1480 | small |
| #4. Normalized Daily Mobility <i>Box-Cox</i> $\lambda=0.1414$ |  mean=0.93 sd=0.19 mean=0.98 sd=0.21 | 28.4030 | <2.2e-16 | 0.2551 | small |
| #5. ROG <i>Box-Cox</i> $\lambda=0.1818$ |  mean=1.31 sd=0.40 mean=1.41 sd=0.59 | 21.6560 | <2.2e-16 | 0.2312 | small |
| #6. Normalized ROG <i>Box-Cox</i> $\lambda=0.1818$ |  mean=0.98 sd=0.32 mean=1.08 sd=0.45 | 28.7070 | <2.2e-16 | 0.2961 | small |

Table 6.1: Statistical analysis of spatial dynamics' features: comparison between migrant and local behaviors.

has been observed. As such, it will assess whether migrants are more prone to visit more municipalities than locals. This feature could be of interest to migration researchers interested in understanding trips to visit the social connections migrants left behind before migration.

2) *Entropy of Visits to Municipalities*. This feature measures the regularity of the visits to different municipalities. Insights from this feature could help future migration research focused on the adaptation of migrants to new settings and routines.

3) *Daily mobility*. This feature is computed using the sequence of BTS towers visited in a day, which is an approximation of the real distance traveled by an individual. I focus on weekday traveled distances to assess the spatial dispersion of migrants potentially due to work-related reasons [189]. As opposed to municipality features that focus on long trips, daily mobility offers a window into short trip behaviors.

4) *Radius of gyration (ROG)*. ROG is computed as the average area covered by of all the BTS towers used by an individual, weighted by the number of calls in each tower. ROG has been used in the literature as an approximation of the distance between home and work, which are the places where people spend most of their time, as has already been shown in the literature using CDR data [7]. We explore this feature as a proxy to evaluate commuting distances, which could motivate future migration researchers to understand the job dispersion that internal migrants are typically exposed to in Mexico. We calculate the ROG for each individual as the root mean square of the average distances from all other locations to his center of

gravity (COG): $ROG_i = \sqrt{\frac{1}{N_i} \sum_{t=1}^{N_i} (d_{it} - COG_i)^2}$.

Finally, given that the spatial dynamics of an individual can be affected by the size and shape of the municipality, I normalize the daily mobility and ROG features by the radius of the municipality where the feature is computed. This process generates two additional spatial dynamics features: 5) Normalized daily mobility and 6) Normalized ROG. I compute all these spatial dynamic features for each in our dataset. Each migrant is characterized by a pre- and a post-migration distribution for each feature and each local is characterized by a unique distribution per feature.

Next, I compute, for each spatial feature, a within-subjects t-test between the pre- and post-migration distributions to assess whether migrants change their spatial dynamics immediately after migration. Most of the tests were statistically significant at $p < 0.05$ *i.e.*, migrants change their spatial behaviors immediately post-migration. In this section, I focus on understanding whether these changes are due to the new physical environment or rather are consequences of the migration process itself. For that purpose, I compare migrants' post-migration behaviors against local behaviors to assess potential spatial dynamic consequences of the migrations. Specifically, I build, for each spatial feature, two population distributions representing the behavior of all locals in our sample and the behavior of all migrants in our sample for that feature. For each feature, I then compare the two distributions using Welch's t-test [190]. Additionally, to quantify the statistical differences between the two populations (effect size) I also compute Cohen's d [191]. The distributions of daily

mobility and ROG can be heavily skewed by a small number of people who have extremely large distances traveled. To avoid the dominance influence of the outliers over the test result, I apply the Box-Cox transformation [192] to the data so that the population distribution is rendered close to normal. Table 6.1 shows the results for the statistical tests. Next, I describe the main observations.

Observation 1. We observe that in the immediate post-migration period, internal migrants in Mexico tend to visit more municipalities and have more irregular behaviors than the local community. The average number of municipalities migrants visit in a month is 2.0, compared to 1.6 for locals. The t-test shows that the difference is significant ($t=34.38$, $p<2.2e-16$), with a small effect size (Cohen's d value is 0.3189). A similar result was observed for the entropy of municipalities visited, with medium effect size. Immediately after migrating, migrants have, on average, higher entropy than locals, showing more irregular mobility patterns. Migrants were observed to visit both their pre-migration municipalities as well as municipalities located close to their post-migration destinations. We hypothesize that these findings could reveal that individuals make an effort to maintain their local connections in their pre-migration municipalities either because of work or personal reasons; in fact, similar findings have been qualitatively reported in other countries [193, 194]. Further qualitative analyses via interviews or surveys would be necessary to confirm or refute our hypothesis for Mexico. Due to the limitations of the temporal range of the data, I am not able to explore whether, in the long term, these behaviors remain more entropic or stabilize to levels similar to the local population.

Observation 2. The Table also shows that, immediately after migrating,

migrants have significantly longer trips than locals (daily average mobility of 1.21 vs. 1.17 after Box-Cox, 3.78km vs. 3.12km before Box-Cox, with $t=16.285$, $p<2.2e-16$). After a Box-Cox transformation to mitigate for outliers, and after spatial normalization to eliminate the role of the municipality size, the significant difference still prevails, with a small Cohen's d effect size. This result highlights that, at least during the first post-migration months, migrants appear to travel longer distances on a daily basis, which could be potentially related to (i) migrants having more daily short trips than locals within a small geographical area or (ii) migrants having a unique, longer than locals, daily trip. We hypothesize that the former could be indicative of a larger informal job market within the migrant community [195]; while the latter could reveal larger job spatial dispersion (jobs are farther away for migrants than for locals), as shown in [196]. Further qualitative studies will be necessary to evaluate both hypotheses for Mexico.

Observation 3. As explained earlier, the ROG represents the geographical area or physical space where individuals spend *a vast majority* of their time. I make use of this variable to characterize the average distance between *home* and *work* [7]. Our analyses show that both ROG and normalized ROG are significantly different with a moderate difference in quantity between the two (small Cohen's d effect size). This result highlights that the differences in daily mobility discussed in *Observation 2* could be potentially due to longer commute trips, at least during the first months post-migration. Similar findings for commuting distances were revealed by Browder *et al.* in an analysis of commuting patterns of internal migrants in Bangkok, Jakarta, and Santiago using survey and interview data [197]. However,

only further qualitative studies in Mexico will be able to clarify this hypothesis. As shown, the proposed framework can be used as a tool to reveal behavioral insights at a large scale that might motivate new qualitative studies for sociologists, geographers or ethnographers. For example, our analyses favored the longer commute trip hypothesis versus the informal market opportunities hypothesis. Understanding the reasons behind this finding, or whether it can also be replicated across countries, would be an interesting research question to analyze.

Social ties We consider the following features to model the online, cell phone-based social ties of migrants and locals:

1) *Number of Contacts per Month*. I compute the average number of monthly contacts the migrant talks to. I define contacts as individuals with whom reciprocal relationships are set up *i.e.*, where two individuals have at least one reciprocated phone call with a duration longer than five seconds. These filters are set up following Onnela's *et al.* work who have argued that reciprocal calls with long duration can be an indication of some work-, family-, leisure- or service- based relationship, while a single call now and then may carry little information [33]. This feature allows us to model the size of a migrant's cellphone-based social network, and compare it against average sizes of local networks.

2) *Number of Calls per Month*. This feature is used to understand the strength of the relationships that the migrant establishes with her cellphone-based social contacts.

3) *Entropy of Contacts*. This feature is computed weighting the cellphone-based social activity with each contact by its call frequency. The objective is to un-

derstand whether the migrant has a predictable communication pattern or whether it is more entropic.

4) *Number of New Contacts per Month*. I use the first three months of CDR data, prior to migration, to build the pre-migration, cellphone-based social network of a migrant. Onnela *et al.* showed that three consecutive months are enough to reconstruct the social network of an individual [33]. Next, during the post-migration period, I identify as new contacts those who share cell phone activity with an individual and who is not present in her pre-migration social network. This feature is critical to evaluate the temporal expansion of the cellphone-based social network in the post-migration period and quantify its growth.

5) To further characterize the cellphone-based social relationships of migrants, I build upon the feature of new contacts and analyze three additional features: 5a) *Ratio of (%) new contacts to the total number of contacts*, 5b) *Number of calls with new contacts*, and 5c) *Ratio of (%) calls with new contacts to all contacts*. Each of these features is computed as monthly average values per individual.

6) To assess the role that place plays in the cellphone-based social network of migrants, I evaluate 6a) *Ratio of (%) local calls* defined as communications with contacts who live in the post-migration municipality with respect to all calls. The objective of this feature is to understand the weight that individuals give to communications with contacts from their previous home (pre-migration municipality) as opposed to their current home location (post-migration municipality); and 6b) *Ratio of (%) local calls with new contacts* with respect to all local calls, to analyze whether migrants mostly develop new cell phone-based contacts locally, in their new

municipality, or in the distance, with their pre-migration municipality.

Similarly to the analysis with spatial dynamics features, I first evaluate whether pre- and post-migration social features are statistically significantly different and then compare post-migration behaviors against locals. The post-migration behaviors were statistically significantly different from their pre-migration behaviors via within-subjects t-tests ($p < 0.05$). We could not confirm significant differences for the new contact features since we do not have data before the pre-migration period to construct the social networks and identify the creation of new contacts in that period. Nevertheless, I also compare these features against the local behavior.

In this section, I focus on analyzing the behavioral differences between migrant and local populations with respect to social network features with the focus of forming potential hypotheses about the social consequences of the migration process. For that purpose, I compute the social features for all locals in our dataset, build the local population distributions for each feature, and compare them against the migrants' post-migration distributions using Welch's t-test and Cohen's d. Table 6.2 shows our results. Next, I discuss our main observations.

Observation 1. Immediately post-migration, migrants communicate with a similar volume of calls than locals, but with a slightly smaller number of contacts. The Table shows that the difference between migrants and locals is statistically significant, with a small Cohen's d effect size for the number of contacts, and negligible for the number of calls. This result shows that, in the large-scale, internal migrants appear to have fewer social connections, and as a result, more frequent communications with their contacts than their local counterpart. In other words,


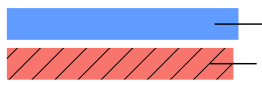

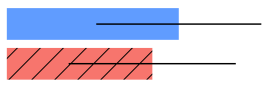
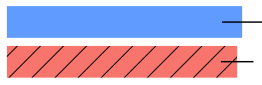
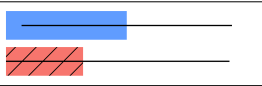

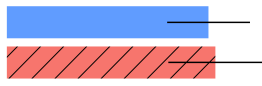
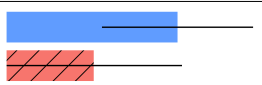
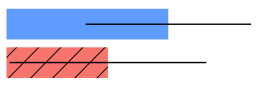
| Features | Mean and SD after Box-Cox  | Welch's T-test | | Cohen's d | |
|--|--|----------------|----------|-----------|-------------|
| | | t | p-value | d | Effect Size |
| #1. Number of Contacts per Month <i>Box-Cox</i> $\lambda = 0.1414$ |  mean=1.31 sd=0.13 mean=1.29 sd=0.13 | -27.7800 | <2.2e-16 | -0.2145 | small |
| #2. Number of Calls per Month <i>Box-Cox</i> $\lambda = -0.1414$ |  mean=1.87 sd=0.26 mean=1.86 sd=0.29 | -4.0602 | 4.92e-5 | -0.0344 | negligible |
| #3. Entropy of Contacts |  mean=1.32 sd=0.63 mean=1.12 sd=0.64 | -40.658 | <2.2e-16 | -0.3205 | small |
| #4. Number of New Contacts per Month <i>Box-Cox</i> $\lambda = -0.1010$ |  mean=1.02 sd=0.09 mean=0.99 sd=0.07 | -29.7430 | <2.2e-16 | -0.2526 | small |
| #5. % New Contacts |  mean=0.11 sd=0.10 mean=0.07 sd=0.13 | -39.107 | <2.2e-16 | -0.3935 | small |
| #6. Number of Calls with New Contacts <i>Box-Cox</i> $\lambda = 0.0202$ |  mean=1.03 sd=0.03 mean=1.04 sd=0.03 | 19.514 | <2.2e-16 | -0.1906 | small |
| #7. % Calls with New Contacts <i>Box-Cox</i> $\lambda = 0.1818$ |  mean=0.60 sd=0.12 mean=0.62 sd=0.14 | 14.2720 | <2.2e-16 | -0.1600 | small |
| #8. % Local Calls |  mean=0.73 sd=0.32 mean=0.37 sd=0.37 | -117.0900 | <2.2e-16 | -1.0871 | large |
| #9. % Local Calls with New Contacts |  mean=0.74 sd=0.38 mean=0.46 sd=0.45 | -43.6890 | <2.2e-16 | -0.7204 | medium |

Table 6.2: Statistical analysis of social ties' features: comparison between migrant and local behaviors using Welch's t-test with Cohen's d and Box-Cox transformation for skewed distributions.

immediately after migration, internal migrants rely on stronger, cell phone-based social relationships than locals.

Observation 2. In the immediate post-migration period, migrants show lower entropy in their cellphone-based social networks than locals (mean=1.12, sd=0.64 vs. mean=1.32, sd=0.63). It highlights the fact that, during their first months post-migration, migrants appear to have a more regular calling behavior. Given that the percentage of local calls is significantly smaller for migrants (see #8 in Table 6.2), we hypothesize that migrants maintain regular calling patterns with their pre-migration municipalities, potentially due to work or family. Similar results have been shown in the context of international migration [198], however further qualitative studies would be necessary to confirm the hypothesis for internal migration in Mexico.

Observation 3. In the first three months after migration, migrants tend to add fewer monthly contacts to their social network than locals. The ratio of new contacts to the existing social network shows a small statistical difference with a mean of 7% for migrants and 11% for locals. This result shows that immediately after migrating, shows significantly lower numbers for communications with local contacts. We hypothesize that this observation might be supported by the theory of acculturation stating that migrants usually have difficulties in their adaption process [145]. While migrants might have the necessity to re-build their disrupted social network, they have difficulties to expand it when compared to non-migrant behaviors. However, further qualitative research would be necessary to confirm such a theory in the context of internal migration in Mexico.

Observation 4. Immediately after migration, migrants have fewer commu-

nications with locals than the locals themselves. While only 37% of the calls that migrants make are with locals, 73% of the calls from locals are to other locals (non-migrants). Similar behavior is observed with the new contacts made. We observe that the ratio of local calls with new contacts is also significantly smaller for migrants; with only 46% of the migrants' calling behavior taking place with new local contacts (as opposed to 74% for locals), which reflects that migrants also continue to expand their social network in their pre-migration communities. These findings reflect that the first months after migration, migrants still heavily rely on their pre-migration social network. We hypothesize that these findings could be explained with the social support theory that determines that migrants tend to seek friends and family social support to buffer the negative effects of migration stress [146]. However, a further qualitative analysis would be necessary to confirm or refute such a theory for Mexican migration.

6.5.2 RQ2: Role of Pre-migration Behaviors on Post-migration Activities

We analyze the role that pre-migration spatial and social features might play in the behavioral changes observed immediately post-migration as migrants adapt to their new communities. For that purpose, I run multivariate regression models on the pre-migration features, with a focus on analyzing (i) the relationship between pre- and post-migration behaviors, (ii) the statistical significance and importance of pre-migration features to the post-migration behaviors, and (iii) to what extent

such models can be used to predict post-migration behaviors given pre-migration information. I compute multivariate regression models with each spatial dynamics and social network feature described in the previous section as both dependent and independent variables. The models are built on a subset of those variables *i.e.*, the independent variables do not include social network features for new contacts since we do not have data before the pre-migration period to construct the social networks and identify the creation of new contacts.

I have tested for multicollinearity between the independent variables used in the regressions via both the variance inflation factor (VIF) and the condition index (CI) [199]. The largest VIF and CI values were for the variable *number of contacts* (with $VIF = 4.3$ and $CI = 12$), while other values were in the range of $1.3 < VIF < 3.5$ and $1 < CI < 7$. Although $VIF < 10$ and $CI < 30$ are considered acceptable [200], I evaluated multivariate regression models with and without the *number of contacts* variable.

Table 6.3 shows the results for each model (one model per line) taking into account the number of contacts. Removing that variable due to potentially low multicollinearity changed the coefficients (size) minimally, but the sign and significance were the same and thus are not reported in the study.

Next, I discuss the main findings.

Observation 1. Post-migration spatial and social features are highly influenced by their pre-migration values *i.e.*, one of the most predictive features for any given post-migration spatial or social feature is its pre-migration value. It is reflected by the significant and positive coefficients (at $p < 0.001$) between the same pre- and

post-migration features, except for the normalized ROG (see bold coefficients in the diagonal of Table 6.3).

However, looking at the value of the regression coefficients, we can observe that the pre-migration features impact their post-migration values differently. For spatial dynamic features (ROG) the impact is small *i.e.*, we have a small rate of change for the post-migration features when its pre-migration values change one unit, and all other pre-migration features are kept the same. On the other hand, the social network features have a larger impact, with rates of change per pre-migration unit having a larger impact on its post-migration values. For example, the radius of gyration before migration (ROG, line #1 in Table) can lead to 0.1713 (***) times change in its post-migration value if the other independent features are kept fixed; while for the number of monthly contacts (line #3) the rate of change in the post-migration values to their pre-migration ones is large: 0.7821 (***). These results also highlight that social network features typically experience larger quantitative changes between the pre- and post-migration stages than the spatial dynamic features, whose changes between pre- and post-migration stages are more modest.

Observation 2. The cellphone-based social relationships that migrants have in the pre-migration period are highly indicative of the types of cell phone-based relationships they have in their immediate post-migration lives. We can observe that the ratio of calls made to the post-migration municipality prior to migrating is significantly and positively related to the post-migration ratio of local calls (#6) and to the post-migration ratio of local calls with new contacts (#7) *i.e.*,

| Post-migration Features | Pre-migration Features | | | | | | | | Adjusted R-squared | p-value |
|----------------------------|------------------------|-------------------|-----------------------|--------------------|------------------------|------------------|------------|---------------|-----------------------|---------|
| | ROG | Normalized ROG | Number of Contacts | Number of Calls | Entropy of Contacts | % Calls to | | Intercept | | |
| | | | | | | Post-migration | Home | | | |
| #1. ROG | 0.1713*** | -0.0378 | 0.9931*** | -0.0107* | -1.7578 | -18.5053*** | 24.3852*** | 0.0281 | <2.2e-16 | |
| #2. Normalized ROG | 0.0334*** | -0.0062 | 0.0992 | 0.0013 | 0.1497 | -6.3999*** | 7.0510*** | <i>0.0120</i> | <2.2e-16 | |
| #3. Number of Contacts | 0.0003 | 0.0003 | 0.7821*** | -0.0012*** | -0.3805*** | -0.2751* | 2.7170*** | 0.4723 | <2.2e-16 | |
| #4. Number of Calls | -0.0365 | -0.0200 | 5.0428*** | 0.5388*** | -27.8281*** | -32.8129*** | 70.9551*** | 0.264 | <2.2e-16 | |
| #5. Entropy of Contacts | 0.0001 | 0.0001 | 0.0240*** | -0.0003*** | 0.4415*** | 0.0450** | 0.4961*** | 0.3516 | <2.2e-16 | |
| #6. % Local Calls (LC) | 0.0003*** | 0.0000 | 0.0009 | -6.854e-5*** | 0.0258** | 0.5942*** | 0.1643*** | 0.2270 | <2.2e-16 | |
| #7. % LC with New Contacts | 0.0003** | 0.0010* | 0.0020 | -0.0001. | 0.0019 | 0.2963*** | 0.3562*** | <i>0.0461</i> | <2.2e-16 | |
| #8. Number of New Contacts | 0.0001 | 0.0003 | 0.0795*** | -0.0004*** | -0.0918** | -0.0564. | 0.4421*** | 0.0926 | <2.2e-16 | |

Significance levels: 0 '***'; 0.001 '**'; 0.01 '*'; 0.05 '.'; 0.1 ''

Table 6.3: Multivariate regressions on pre-migration behavioral features to quantify their role on post-migration behaviors. Adjusted R-squared values are used to assess the predictability of post-migration behaviors with pre-migration features.

migrants who start developing new cellphone-based social relationships with their post-migration municipalities even before they migrate, are prone to have more local communications and more new local contacts once they move to their post-migration destination.

Similarly, the number of contacts pre-migration is also significantly related (0.0697 at significance level 0.05) to the number of new contacts post-migration (#8) *i.e.*, people who are more socially connected prior to migrating will highly probably develop a large social network with new contacts in their post-migration communities, immediately after migrating. Observations 1 and 2 could potentially motivate further qualitative studies to assess the theory that migrants' personality characteristics can partially explain their adaptability in the host society [145]. Although it is not enough to fully explain how they cope with their new environment, previous work has shown that people who communicate more are generally better at re-rooting their social network in the host society [124].

Observation 3. The pre-migration ratio of calls made to the post-migration municipality is statistically significantly related to both the ROG (#1) and the normalized ROG (#2). We observe a significant and high negative coefficient for the ROG (-18.5053 ***) and the normalized ROG (-6.3999 ***). This result highlights that having social connections in the pre-migration period with the post-migration destination affect the estimated rate of change for both spatial dynamics variables by a high negative factor when all other features are kept fixed. In other words, the more social connections during pre-migration with the post-migration municipality, the more significant the decrease in the average length of the commuting patterns in the immediate post-migration period.

This result is interesting when combined with a result obtained in the previous Section 6.5.1. The analysis comparing post-migration behaviors between migrants and locals showed that, on average, migrants tend to have larger ROGs and normalized ROGs. The current analysis appears to be showing that, if the migrant already has a connection with the post-migration destination, the ROG will be smaller, and potentially closer to that of locals. Further qualitative studies could look into whether this result could be interpreted as an indication of migrants adapting more quickly to their new environments when they previously have social connections with and knowledge about those environments. Similar results have been shown by Holmes *et al.* who interviewed migrants and assessed that communities with access to information from their post-migration communities prior to migration helped in bridging the challenges of adaptation [201].

Observation 4. The adjusted R-squared values shown in the Table mea-

sure how much variance of the dependent variables (the post-migration behavioral features) can be explained by pre-migration features. Generally, these values are good for some of the social network features ($0.264 < R^2 < 0.4723$), and poor for the spatial dynamics features (maximum adjusted $R^2 = 0.0281$). Specifically, the regression models for the number of contacts, the entropy of contacts and the ratio of local calls are relatively high, showing that these post-migration behaviors can be partially predicted with the subset of pre-migration features considered. However, the models for ROG, normalized ROG and the ratio of new contacts have small R-squared values, revealing that we can not use only pre-migration features to predict these post-migration values.

6.6 Potential Limitations

At the time when the CDR data were processed (2009 – 2010), it is estimated that the percentage of cell phone owners in Mexico was approximately 60% [202]. Additionally, given that we only have access to CDR data from one cell phone company, this limits the representativeness of the local and migrant populations. Previous work has shown that cell phone usage has a high penetration rates in Latin American countries. The data cover a population with different demographic and socioeconomic characteristics, with percentages of cell phone users per strata similar to the actual population census [26].

However, in this study, there is yet another type of selection bias. Since home location is inferred from CDR data, individuals that do not have call activity

at night, are not assigned a home location and, as a result, are not part of the behavioral analysis. This approach implies that the behavior of people who have an overall reduced number of communications will probably be filtered out from the analyses; which might translate into losing information from a specific migrant group associated to low cell phone use.

A second limitation of the work presented in this study is the temporal length of the dataset used, which is of eight months. Since I focus the analyses on immediate post-migration behaviors only, in principle, I do not require long periods of CDR data after migration. However, due to the limited temporal length, the approach is not able to differentiate between long-term and short-term migrants. In any case, the methods and results presented in this study are still highly valuable to understand immediate post-migration behaviors across typologies of internal migrants.

A third limitation of the study is the CDR data we used. Although the data can reveal spatial and cell phone-based social behaviors of migrants, it lacks insight into the reasons why the observed behaviors take place including, but not limited to, the psychological and decision-making processes related to the migration experience.

Finally, it is important to highlight that our analyses use data subject to privacy concerns. Although the data are anonymized and the results are reported in an aggregated manner, it has been previously shown that under some circumstances CDR data can be potentially used to identify individuals [203]. Another ethical concern is that the methods for the inference of migrants might be used by unintended people and expose migrants to potential risks. However, the methods we developed are highly adapted to certain types of data, which are not validated on

other types of digital traces. If the data are well managed under user privacy policy, the methods alone present little risks.

Chapter 7: Conclusions and Future Directions

7.1 Conclusions

To conclude, in the three studies, I create data-driven models that sense, characterize, and predict the patterns of mobility activities, social interactions, and communications, for people under different socioeconomic context or facing natural disasters. The common objective is to enhance decision-making processes with large-scale digital traces in the area of Smart and Connected Communities. In the following sections, I summarize the technical contributions I have made in the three studies. I also present the potential societal impacts and how these methods can be applied in decision-making processes.

7.1.1 Topic Models to Infer Socioeconomic Maps

In the study of inferring socioeconomic maps with cell phone data, I proposed a framework that incorporates models that can infer the socioeconomic level of regions based on extracted behavior features from digital traces. Specifically, I proposed an innovative topic modeling method to extract population mobility patterns from the spatiotemporal trajectory of individual users. This method has proved to increase

the state-of-art prediction performance by 9%. Better models with higher prediction accuracy can lead to more reliable estimation. Therefore it is important to consider the improvement of data-driven models.

One may argue that if the inferred socioeconomic index is already done by household surveys or censuses, there is little gain by using CDRs to infer the existing survey data. The point is that CDRs-based data-driven method is not a substitute for the existing system, but a complement to current survey methods. In terms of societal impact, there are many practical conditions this method can be applied.

Reliable survey data is not always available or up to date, since household surveys cost tens to hundreds of millions of dollars to collect [50]. Many developing countries especially lack complete and timely estimates for antipoverty programs and resource allocation [151]. CDRs-based methods help to fill the gaps where or when survey data are not available. The estimated data is also of great value although the accuracy may not be 100% [14]. Next, I describe three important contributions of this study:

Provide estimates where no information is available. Cell phones are widely used in developing countries. Digital traces may exist in areas where reliable survey data is not available. Based on behavior patterns extracted from digital traces, we can infer the socioeconomic estimates for these areas. In countries with political instabilities or turmoils where no recent survey data is available, it is possible to estimate the socioeconomic statuses based on CDRs.

Generate maps with high spatial resolution. Household surveys with a limited sample size usually lead to SEL maps with coarse spatial resolution, for

example at the state level. However, aid programs or policies may require more detailed information at higher spatial resolution level, such as the county or neighborhood level. With CDRs, it is possible to achieve estimates at tower coverage level, which is usually less than 1 km² in dense urban areas or varies with more than 4 km² in rural areas. However, we still need to consider the possible effect by space and geography on model accuracy.

Generate up-to-date or future estimates. Surveys are usually conducted every several years due to the high cost. While CDRs are continuously generated by users when they use cell phones. CDRs can be used to update SEL maps, providing current information for decision-making. It is also possible to model how population behaviors might relate to future socioeconomic change. This model is not discussed in the study due to the limitation of data for the experiment. The idea is that CDRs-based data-driven methods enable up-to-date estimates or even future estimates for decision-making.

The data-driven model proposed in this study has been applied to an interactive visualization tool for the World Bank, which shows the predicted results of poverty rates in Guatemala at municipality level [15].

7.1.2 Understanding Online Communications of Citizens and Local Governments during Natural Disasters

This study proposes a semi-automatic framework to extract and compare the communications of citizens and local governments under different conditions. The

framework requires few human efforts in collecting and analyzing tweets. It uses rank by cross-entropy and a fully unsupervised ST-LDA model to select disaster-related tweets and extract specific topics that are discussed online during natural disasters. Although it requires human interpretation of topics, it minimizes the involvement of human beings in information retrieval and summarizing.

In this study, I show three examples of how the topic labeled digital traces can be used. First, I use spatiotemporal analysis to identify spatiotemporal clusters with bursts of tweets. Topics of these clusters reflect the issues that residents are mostly concern about, which can be used to identify the priority of their needs. In the identified spatial areas, we find differences between user needs and the issues local governments have communicated. I also analyze topics and themes on different types of roads, and how residents interact with local government accounts for different issues. The findings might help local governments to react accordingly to residents' concerns, and to adjust communication strategies in disasters.

The applied scenarios are not limited to what I have listed here. With the labeled disaster-related tweets, we are able to identify the priority of needs communicated by residents in different communities. Communities here refer to geographically delineated areas at different levels, for example, the streets, neighborhoods or counties. By changing geographies, decision makers will be able to explore the types of needs associated to different communities. Additionally, as the communication themes and topics might change before and after disasters, these might reflect priority changes at different stages of the disaster, offering information for disaster preparedness, response, and recovery.

The framework is not only useful for understanding local specific issues during snowstorms but also applicable to other disasters since the framework doesn't incorporate any prior knowledge about the disaster in the model. In this study, I use data of snowstorms as an example to show what issues are communicated during snowstorms. The framework is applicable to other emergency situations, where we know many residents are affected, but do not have the real-time information regarding where and how they are affected, for example hurricanes and earthquakes. The framework helps to retrieve and summarize the issues residents claim in online communications.

7.1.3 Identification and Characterization of Internal Migrants with Cell Phone Data

In the migration study, I present a framework that can be used to understand migrants' behaviors with cell phone traces. The internal migrants are identified based on their home location change, which is learned with spatiotemporal trajectories extracted from cell phone data. The proposed framework enables to examine the difference between migrants and local people from perspectives of spatial dynamics and social ties, and assesses determinants for the adaptation of migrants to the host society.

This study shows digital traces such as CDRs enable examination of micro-level behaviors at a large scale. In the study, I show some interesting observations which have not been examined before. For example, the mobility behaviors and

social relations of migrants are usually not studied together at a large scale. One of the findings is that in the immediate post-migration period, migrants tend to have longer mobility distances but fewer local contacts compared to local people. It seems that migrants try to make up the disadvantage in social capital with mobility behaviors. This could be an interesting hypothesis for sociological researchers to look into. The study enables us to have a better understanding of migrants' behaviors. It also helps to identify the issues that tend to be most important for the adaptation of migrants, i.e. social relations, therefore enable informed policy-making and aid efforts for decision makers.

This study also shows that digital traces can facilitate studies on the evaluation of effects on research subjects with intervening conditions. In the traditional research methods [54], the study on effect evaluation relies on laboratory experiments or experimental simulations, where control is emphasized to understand the status of subjects before and after the intervention. With digital traces, it becomes much easier to observe subjects under multiple control and stimulus conditions [47]. Due to the characteristics of 'continuous in time', it is possible for researchers to identify the subjects that meet certain pre-defined experimental conditions and observe the change to evaluate the short-term or long-term effect. For examples, digital traces can be used to evaluate the short-term or long-term behavior changes after disasters.

7.2 Limitations

Large-scale data sources have tremendous potential to make better behavioral predictions, enhance situation awareness, and create knowledge, therefore leading to overall increased quality of life. Although data has great value, there are limitations to be considered and presented from an ethical perspective. In the previous chapters, I have listed the limitations specific to each study based on the type of data used and the studied question. Here I present general ethical issues in data-driven research, which might be helpful for researchers to regularize their way of research to respect the principles of ethics. The criticisms to data-driven research for decision making mostly focus on two aspects: the biased representation of data and the potential harms to personal privacy.

Large-scale digital traces reflect the status of a much larger population than traditional survey methods, however, there are inevitable biases in the data regarding the representativeness of all populations. The biases can be induced by both user and data provider sides. Digital traces are details of the user's interactive behaviors with digital devices. Accessibility to digital devices is a necessary condition to have digital traces that represent one's behavior. However, accessibility can be hindered due to economic reasons or physiological restrictions [204].

It is critical to identify the biases in data representation, since a unified statistical model applied to the data regardless of the actual human behind the data may lead to biased conclusions. If using the conclusions as they represent the whole populations, policies based on these will also be biased. Biased policies or actions

that only aim for the goods of certain groups of people might lead to decisions that are detrimental to the overall fairness of a society.

For example, in my previous work, I have compared several techniques that are used for identification of migrants using cell phone traces. The work evaluated the accuracy of detection for different types of rural and urban migrants based on ground truth that is extracted from census data. Results showed the accuracy is relatively low for people migrating from or to rural areas. The study also enables us to quantify the amount of bias that is induced with data-driven methods [205].

Digital traces are personal data, which present potential risks to personal privacy. One of the potential risks is that digital traces might be used to reveal the identity or other private information, which is also commonly known as personal identifiable information (PII), by linking data from different sources [206]. Human activities are highly individualized and unique. Even information that directly related to identities are eliminated or encrypted, one's identity can still be inferred based on the unique behaviors revealed in the digital traces [203]. Researchers may not misuse digital traces, but data breach might present harm to users. The vulnerability in data storage can lead data intruders to access data and use it in harmful ways [207]. Researchers need to consider security issues in data storage and data transfer.

Digital traces reveal information such as home and work locations [208], migration status [209], psychological status [210, 211], or health issues [212] that is unintended by users. Individual mobility behaviors present high regularity that is dependent on the location of homes and workplaces [7], which means that with

enough mobility trajectories, it is possible to infer the home and work locations of users. There are also behavioral clues that indicate psychological or health status from digital traces, which enable algorithms to detect certain groups of people. However, users may not expect this kind of information to be inferred and exposed to the public. To ease these problems, we need to assess the potential risks in each step of work and adjust policies in data and results publication accordingly. For example, user ids should be cautiously referred to in publications depending on whether it is public figure account, organizational account or private individual account. In this work, to follow ethical norms regarding privacy, encrypted ids have been used, and no personal information has been accessed or released in any of the publications.

7.3 Future Directions

Next, I describe potential extensions to the work I have presented in this dissertation. The methods I have presented in this dissertation are not only applicable to the specific decision-making scenarios I have explored, but rather, they allow to extract regular behavior information from any type of discrete digital traces and to measure the relations between behavior change and the environmental context. In Section [7.3.1](#), I describe other empirical studies that could be done using the proposed methods. On the other hand, in Section [7.3.2](#) I will present potential directions to enhance the computational methods presented in this dissertation. One line of work is to explore methods for more accurate and multi-dimensional behavior modeling. It is also critical to think about the generalization of data-driven

models: considering factors that might impact the performance of models, how to maintain the effectiveness of these models for decision-making processes, and how to incorporate bias analyses in the design of data-driven models.

7.3.1 Validation in Other Scenarios

In Chapter 4, I present a topic modeling method to extract population mobility patterns for the prediction of socioeconomic levels. Mobility behaviors are not only related to socioeconomic change, but also related to the labor force in the market [213], societal characteristics of regional segmentation [214], and crime rate [215]. It would be interesting to explore whether mobility patterns can be used to infer these contextual indexes, what kind of population mobility patterns are indicators, and whether these patterns have been explained comprehensively in existing theories. On one hand, this kind of empirical experiments would help to expand the application areas of the data-driven methods, on the other hand, it can potentially provide empirical evidence or supplement the existing theories in these fields.

In Chapter 5, I describe a semi-automatic framework to understand the online communications of citizens and local governments. The framework is mainly used to retrieve disaster-related tweets and summarize the topics. There is no prior information about disasters. Therefore, this framework can be applied to other disaster scenarios such as earthquakes, hurricanes, or floods. Another potential application is to apply the framework on data that comes from several disasters of the same type and compare the difference of user behaviors combined with government

response to the disasters. This type of comparative analyses can help decision-makers to evaluate the effectiveness of policies in disaster response.

Chapter 6 presents a data-driven framework for the study of internal migrants with cell phone traces. The proposed method can be applied to other types of data which contain continuous input of geographical information from users, for example, the geotagged social media data or email logs. Geotagged tweets and email logs have been used to study internal and international migration flows [32], but few studies have looked into the individual-level behaviors of migrants. Cell phone data is usually unavailable for international calls due to roaming. While social media data are usually generated by one service provider with a similar metadata format worldwide. Therefore, social media data can extend the study of migrants to the global level.

7.3.2 Development of Modeling Methods

There are potential directions to extend this work for future studies.

Generalization of models. The performance of prediction models in this dissertation are tested multiple times with test data sets. However, there is still the question of whether models work well for other geographic areas or in other time frames. We need to consider the factors that might affect the scalability of the model in geographical and time scales, to test the generalization of models. Given the availability of data, it might be interesting to apply the models on other data sets, for example, data from other countries.

As user behaviors may change with the emergence of new technologies, models built with pre-determined features may not work well in the future. For example, instant messaging are replacing part of the phone calls for communication for people who can afford smartphones and wireless networking. The relations between phone call behaviors and socioeconomic levels might change. It would be interesting to think about the valid life of a model, what the conditions are for the model to continue working with the claimed accuracy, and how the models might be adapted to reflect the changes.

Granular characterization of communication behaviors. In Chapter 5, I used a framework to extract themes and topics of communication. However, for one topic users may have different motivations to post, and share different perspectives of information. For example, among the tweets that are labeled with the topic *snow removal*, some report issues as where the snow needs to be removed; some compliment local governments work; and some share the information that snow has been removed. Themes and topics show what issues citizens mostly care about. A better understanding of how they discuss these issues would help local governments to better understand how to respond. To have a fine granular understanding of the communications, we may need to use techniques that consider language features more than n-gram of words, for example, the syntactic notation, or create more linguistically refined models. This can be a potential direction to improve modeling of online communications.

Interpretation of behaviors. In Chapter 4 and Chapter 6, I explored methods to model physical behaviors such as social activities and mobility with cell phone

traces. But cell phone data usually do not contain the content of communications, which limits us to further understand conversations and emotions that are associated with their behaviors. In Chapter 5, although I analyzed topics and themes of communication, I haven't looked into users' social or spatial behaviors. Geotagged tweets also reflect mobility trajectory of users. With geotagged tweets, we might look into the mobility behaviors of users and analyze the conversations associated with mobility behaviors. It may help us to understand not only how people move in disasters but also why since conversations usually contain the information that describes environments or express feelings at a certain location. Previous works have used social media data to infer the meaning of places to explain trajectories [216], or used probabilistic models to discover the location meanings [217], but seldom has studied the purpose of mobility or their psychological changes during mobility. Geotagged social media data might also be used to detect migrants. From the text, we might have more subjective information regarding their migration.

Bias in data-driven models In Section 7.2 I presented the bias and privacy issues in data-driven models. The 'behavioral' digital traces are usually large but not equally representative of the different groups of people due to user accessibility and user preferences. It is important to understand which groups of people are underrepresented, where the data bias comes from, and how to measure such bias. To this end, we can design weighted mechanisms or build probabilistic models that incorporate such information to reduce bias. We may also consider designing data-driven models that have more explainable components to enhance the transparency of models, which might enable the decision-makers to be aware of the potential bias.

Bibliography

- [1] Jon D Mcauliffe and David M Blei. Supervised topic models. In *Advances in neural information processing systems*, pages 121–128, 2008.
- [2] International Telecommunication Union. *The world in 2014: ICT facts and figures*. ITU, 2014.
- [3] Lingzi Hong, Cheng Fu, Paul Torrens, and Vanessa Frias-Martinez. Understanding citizens’ and local governments’ digital communications during natural disasters: The case of snowstorms. In *Proceedings of the 2017 ACM on Web Science Conference*, pages 141–150. ACM, 2017.
- [4] Brian J Meacham. Integrating human behavior and response issues into fire safety management of facilities. *Facilities*, 17(9/10):303–312, 1999.
- [5] Charles Zastrow and Karen Kirst-Ashman. *Understanding human behavior and the social environment*. Cengage Learning, 2006.
- [6] Edward T Jennings Jr and Jeremy L Hall. Evidence-based practice and the use of information in state agency decision making. *Journal of Public Administration Research and Theory*, 22(2):245–266, 2011.
- [7] M. Gonzalez, C. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *nature*, 453(7196):779, 2008.
- [8] Rumki Basu. *Public administration: Concepts and theories*. Sterling Publishers Pvt. Ltd, 2004.
- [9] Junghoon Chae, Dennis Thom, Harald Bosch, Yun Jang, Ross Maciejewski, David S Ebert, and Thomas Ertl. Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition. In *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on*, pages 143–152. IEEE, 2012.
- [10] Johann Höchtl, Peter Parycek, and Ralph Schöllhammer. Big data in the policy cycle: Policy decision making in the digital era. *Journal of Organizational Computing and Electronic Commerce*, 26(1-2):147–169, 2016.

- [11] Eiman Al Nuaimi, Hind Al Neyadi, Nader Mohamed, and Jameela Al-Jaroodi. Applications of big data to smart cities. *Journal of Internet Services and Applications*, 6(1):25, 2015.
- [12] National Science Foundation. Smart and connected communities (s&cc) program solicitation. <https://www.nsf.gov/pubs/2019/nsf19564/nsf19564.pdf>, 2019. Accessed: 2019-04-10.
- [13] Amanda Clarke and Helen Margetts. Governments and citizens getting to know each other? open, closed, and big data in public management reform. *Policy & Internet*, 6(4):393–417, 2014.
- [14] Matthew J Salganik. *Bit by bit: social research in the digital age*. Princeton University Press, 2017.
- [15] Marco Hernandez, Lingzi Hong, Vanessa Frias-Martinez, and Enrique Frias-Martinez. *Estimating poverty using cell phone data: evidence from Guatemala*. The World Bank, 2017.
- [16] Bevaola Kusumasari, Quamrul Alam, and Kamal Siddiqui. Resource capability for local government in managing disaster. *Disaster Prevention and Management: An International Journal*, 19(4):438–451, 2010.
- [17] Sonja Utz, Friederike Schultz, and Sandra Glocka. Crisis communication online: How medium, crisis type and emotions affected public reactions in the fukushima daiichi nuclear disaster. *Public Relations Review*, 39(1):40–46, 2013.
- [18] Rosaria Conte, Nigel Gilbert, Giulia Bonelli, Claudio Cioffi-Revilla, Guillaume Deffuant, Janos Kertesz, Vittorio Loreto, Suzy Moat, J-P Nadal, Anxo Sanchez, et al. Manifesto of computational social science. *The European Physical Journal Special Topics*, 214(1):325–346, 2012.
- [19] Kay W Axhausen. Social networks, mobility biographies, and travel: survey challenges. *Environment and Planning B: Planning and design*, 35(6):981–996, 2008.
- [20] Alessandro Vespignani. Predicting the behavior of techno-social systems. *Science*, 325(5939):425–428, 2009.
- [21] Santi Phithakkitnukoon, Zbigniew Smoreda, and Patrick Olivier. Socio-geography of human mobility: A study using longitudinal mobile phone data. *PloS one*, 7(6):e39253, 2012.
- [22] Sandra González-Bailón. Social science in the era of big data. *Policy & Internet*, 5(2):147–160, 2013.
- [23] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. AcM, 2010.

- [24] Vincent D Blondel, Markus Esch, Connie Chan, Fabrice Clérot, Pierre Deville, Etienne Huens, Frédéric Morlot, Zbigniew Smoreda, and Cezary Ziemlicki. Data for development: the d4d challenge on mobile phone data. *arXiv preprint arXiv:1210.0137*, 2012.
- [25] Kevin Makice. *Twitter API: Up and running: Learn how to build applications with the Twitter API.* ” O’Reilly Media, Inc.”, 2009.
- [26] Vanessa Frias-Martinez and Jesus Virseda. Cell phone analytics: Scaling human behavior studies into the millions. *Information Technologies & International Development*, 9(2):pp–35, 2013.
- [27] Morgan R Frank, Lewis Mitchell, Peter Sheridan Dodds, and Christopher M Danforth. Happiness and the patterns of life: A study of geolocated tweets. *Scientific reports*, 3:2625, 2013.
- [28] Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H Chi. Tweets from justin bieber’s heart: the dynamics of the location field in user profiles. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 237–246. ACM, 2011.
- [29] Sibren Isaacman, Richard Becker, Ramón Cáceres, Margaret Martonosi, James Rowland, Alexander Varshavsky, and Walter Willinger. Human mobility modeling at metropolitan scales. In *Proceedings of the 10th international conference on Mobile systems, applications, and services*, pages 239–252. Acm, 2012.
- [30] Kevin S Kung, Kael Greco, Stanislav Sobolevsky, and Carlo Ratti. Exploring universal patterns in human home-work commuting from mobile phone data. *PloS one*, 9(6):e96180, 2014.
- [31] Justine I Blanford, Zhuojie Huang, Alexander Savelyev, and Alan M MacEachren. Geo-located tweets. enhancing mobility maps and capturing cross-border movement. *PloS one*, 10(6):e0129202, 2015.
- [32] Emilio Zagheni, Venkata Rama Kiran Garimella, Ingmar Weber, et al. Inferring international and internal migration patterns from twitter data. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 439–444. ACM, 2014.
- [33] J-P Onnela, Jari Saramäki, Jorkki Hyvönen, György Szabó, David Lazer, Kimmo Kaski, János Kertész, and A-L Barabási. Structure and tie strengths in mobile communication networks. *Proc. Nat. Academy of Sciences*, 104(18):7332–7336, 2007.
- [34] Riitta Toivonen, Jussi M Kumpula, Jari Saramäki, Jukka-Pekka Onnela, János Kertész, and Kimmo Kaski. The role of edge weights in social networks: modelling structure and dynamics. In *Noise and Stochastics in Complex Systems*

and Finance, volume 6601, page 66010B. International Society for Optics and Photonics, 2007.

- [35] Ai-Ju Huang, Hao-Chuan Wang, and Chien Wen Yuan. De-virtualizing social events: understanding the gap between online and offline participation for event invitations. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 436–448. ACM, 2014.
- [36] Nan Lin, Karen S Cook, and Ronald S Burt. *Social capital: Theory and research*. Transaction Publishers, 2001.
- [37] Martha McCaughey and Michael D Ayers. *Cyberactivism: Online activism in theory and practice*. Routledge, 2013.
- [38] Neil Dufty et al. Using social media to build community disaster resilience. *Australian Journal of Emergency Management, The*, 27(1):40, 2012.
- [39] Xingjie Liu, Qi He, Yuanyuan Tian, Wang-Chien Lee, John McPherson, and Jiawei Han. Event-based social networks: linking the online and offline social worlds. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1032–1040. ACM, 2012.
- [40] Jey Han Lau, Nigel Collier, and Timothy Baldwin. On-line trend analysis with topic models: \# twitter trends detection topic model online. *Proceedings of COLING 2012*, pages 1519–1534, 2012.
- [41] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. Comparing twitter and traditional media using topic models. In *European conference on information retrieval*, pages 338–349. Springer, 2011.
- [42] Kentaro Sasaki, Tomohiro Yoshikawa, and Takeshi Furuhashi. Online topic model for twitter considering dynamics of user interests and topic trends. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1977–1985, 2014.
- [43] Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. Target-dependent twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 151–160. Association for Computational Linguistics, 2011.
- [44] Johan Bollen, Huina Mao, and Alberto Pepe. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *Fifth International AAAI Conference on Weblogs and Social Media*, 2011.
- [45] Yla R Tausczik and James W Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54, 2010.

- [46] Mattia Samory and Tanushree Mitra. 'the government spies using our webcams': The language of conspiracy theories in online discussions. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):152, 2018.
- [47] Ray M Chang, Robert J Kauffman, and YoungOk Kwon. Understanding the paradigm shift to computational social science in the presence of big data. *Decision Support Systems*, 63:67–80, 2014.
- [48] Lisa F Berkman and S Leonard Syme. Social networks, host resistance, and mortality: a nine-year follow-up study of alameda county residents. *American journal of Epidemiology*, 109(2):186–204, 1979.
- [49] Newzoo. Global mobile market report. <https://newzoo.com/insights/rankings/top-50-countries-by-smartphone-penetration-and-users/>, 2018. Accessed: 2019-02-26.
- [50] Joshua Evan Blumenstock. Fighting poverty with data. *Science*, 353(6301):753–754, 2016.
- [51] Kathleen J Tierney. From the margins to the mainstream? disaster research at the crossroads. *Annu. Rev. Sociol.*, 33:503–525, 2007.
- [52] Felix Ritchie. Resistance to change in government: risk, inertia and incentives. *Working Paper*, 2014.
- [53] Kenneth A Lachlan, Patric R Spence, and Xialing Lin. Expressions of risk awareness and concern through twitter: On the utility of using the medium as an indication of audience needs. *Computers in Human Behavior*, 35:554–559, 2014.
- [54] Joseph E McGrath. Dilemmatics: The study of research choices and dilemmas. *American Behavioral Scientist*, 25(2):179–210, 1981.
- [55] Robert G D'Eon, Robert Serrouya, Graham Smith, and Christopher O Kochanny. Gps radiotelemetry error and bias in mountainous terrain. *Wildlife Society Bulletin*, pages 430–439, 2002.
- [56] Daniele Quercia, Licia Capra, and Jon Crowcroft. The social world of twitter: Topics, geography, and emotions. In *Sixth International AAAI Conference on Weblogs and Social Media*, 2012.
- [57] Saint John Walker. Big data: A revolution that will transform how we live, work, and think, 2014.
- [58] Muhammad Imran, Shady Elbassuoni, Carlos Castillo, Fernando Diaz, and Patrick Meier. Practical extraction of disaster-relevant information from social media. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 1021–1024. ACM, 2013.

- [59] Lingzi Hong, Weiwei Yang, Philip Resnik, and Vanessa Frias-Martinez. Uncovering topic dynamics of social media and news: The case of ferguson. In *International Conference on Social Informatics*, pages 240–256. Springer, 2016.
- [60] Dirk Helbing and Stefano Balietti. From social data mining to forecasting socio-economic crises. *The European Physical Journal Special Topics*, 195(1):3, 2011.
- [61] James P Bagrow and Yu-Ru Lin. Mesoscopic structure and social aspects of human mobility. *PloS one*, 7(5):e37676, 2012.
- [62] Jing Yuan, Yu Zheng, and Xing Xie. Discovering regions of different functions in a city using human mobility and pois. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 186–194. ACM, 2012.
- [63] Justin Cranshaw, Raz Schwartz, Jason Hong, and Norman Sadeh. The livehoods project: Utilizing social media to understand the dynamics of a city. In *Sixth International AAI Conference on Weblogs and Social Media*, 2012.
- [64] Fei Wu, Zhenhui Li, Wang-Chien Lee, Hongjian Wang, and Zhuojie Huang. Semantic annotation of mobility data using social media. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1253–1263. ACM, 2015.
- [65] Carmen Karina Vaca, Daniele Quercia, Francesco Bonchi, and Piero Fraternali. Taxonomy-based discovery and annotation of functional areas in the city. In *Ninth International AAI Conference on Web and Social Media*, 2015.
- [66] Jiahui Wu, Lingzi Hong, and Vanessa Frias-Martinez. Predicting perceived cycling safety levels using open and crowdsourced data. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 1669–1676. IEEE, 2018.
- [67] Michael Xie, Neal Jean, Marshall Burke, David Lobell, and Stefano Ermon. Transfer learning from deep features for remote sensing and poverty mapping. *arXiv preprint arXiv:1510.00098*, 2015.
- [68] Neal Jean, Marshall Burke, Michael Xie, W Matthew Davis, David B Lobell, and Stefano Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, 2016.
- [69] Venkata Sasank Pagolu, Kamal Nayan Reddy, Ganapati Panda, and Babita Majhi. Sentiment analysis of twitter data for predicting stock market movements. In *2016 international conference on signal processing, communication, power and embedded system (SCOPEs)*, pages 1345–1350. IEEE, 2016.
- [70] Brendan O’Connor, Ramnath Balasubramanyan, Bryan R Routledge, Noah A Smith, et al. From tweets to polls: Linking text sentiment to public opinion time series. *Icwsn*, 11(122-129):1–2, 2010.

- [71] Alberto Rubio, Vanessa Frias-Martinez, Enrique Frias-Martinez, and Nuria Oliver. Human mobility in advanced and developing economies: A comparative analysis. In *AAAI Spring Symposium: Artificial Intelligence for Development*, 2010.
- [72] Nathan Eagle, Michael Macy, and Rob Claxton. Network diversity and economic development. *Science*, 328(5981):1029–1031, 2010.
- [73] Vanessa Frias-Martinez, Cristina Soguero-Ruiz, Enrique Frias-Martinez, and Malvina Josephidou. Forecasting socioeconomic trends with cell phone records. In *Proceedings of the 3rd ACM Symposium on Computing for Development*, page 15. ACM, 2013.
- [74] Christopher Njuguna and Patrick McSharry. Constructing spatiotemporal poverty indices from big data. *Journal of Business Research*, 70:318–327, 2017.
- [75] Jameson L Toole, Yu-Ru Lin, Erich Muehlegger, Daniel Shoag, Marta C González, and David Lazer. Tracking employment shocks using mobile phone data. *Journal of The Royal Society Interface*, 12(107):20150185, 2015.
- [76] Abdullah Almaatouq, Francisco Prieto-Castrillo, and Alex Pentland. Mobile communication signatures of unemployment. In *International conference on social informatics*, pages 407–418. Springer, 2016.
- [77] Dolan Antenucci, Michael Cafarella, Margaret Levenstein, Christopher Ré, and Matthew D Shapiro. Using social media to measure labor market flows. Technical report, National Bureau of Economic Research, 2014.
- [78] Linna Li, Michael F Goodchild, and Bo Xu. Spatial, temporal, and socioeconomic patterns in the use of twitter and flickr. *cartography and geographic information science*, 40(2):61–77, 2013.
- [79] Daniele Quercia, Diarmuid Ó Séaghdha, and Jon Crowcroft. Talk of the city: Our tweets, our community happiness. In *ICWSM*, 2012.
- [80] Joshua Blumenstock and Nathan Eagle. Mobile divides: gender, socioeconomic status, and mobile phone use in rwanda. In *Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development*, page 6. ACM, 2010.
- [81] Joshua E Blumenstock. Calling for better measurement: Estimating an individuals wealth and well-being from mobile phone transaction records. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, Workshop on Data Science for Social Good*, 2015.

- [82] Christopher Smith-Clarke, Afra Mashhadi, and Licia Capra. Poverty on the cheap: Estimating poverty maps using aggregated mobile communication networks. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 511–520. ACM, 2014.
- [83] Christopher Smith, Afra Mashhadi, and Licia Capra. Ubiquitous sensing for mapping poverty in developing countries. *Paper submitted to the Orange D4D Challenge*, 2013.
- [84] Victor Soto, Vanessa Frias-Martinez, Jesus Virseda, and Enrique Frias-Martinez. Prediction of socioeconomic levels using cell phone records. In *International Conference on User Modeling, Adaptation, and Personalization*, pages 377–388. Springer, 2011.
- [85] Thoralf Gutierrez, Gautier Krings, and Vincent D Blondel. Evaluating socio-economic state of a country analyzing airtime credit and mobile phone datasets. *arXiv preprint arXiv:1309.4496*, 2013.
- [86] Abdisalan M Noor, Victor A Alegana, Peter W Gething, Andrew J Tatem, and Robert W Snow. Using remotely sensed night-time light as a proxy for poverty in africa. *Population Health Metrics*, 6(1):5, 2008.
- [87] Muneo Kaigo. Social media usage during disasters and social capital: Twitter and the great east japan earthquake. *Keio Communication Review*, 34(1):19–35, 2012.
- [88] JooHo Kim and Makarand Hastak. Social network analysis: Characteristics of online social networks after a disaster. *International Journal of Information Management*, 38(1):86–96, 2018.
- [89] Joao Porto De Albuquerque, Benjamin Herfort, Alexander Brenning, and Alexander Zipf. A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management. *International Journal of Geographical Information Science*, 29(4):667–689, 2015.
- [90] Leysia Palen, Starr Roxanne Hiltz, and Sophia B Liu. Online forums supporting grassroots participation in emergency preparedness and response. *Communications of the ACM*, 50(3):54–58, 2007.
- [91] David E Alexander. Social media in disaster risk reduction and crisis management. *Science and engineering ethics*, 20(3):717–733, 2014.
- [92] Leysia Palen and Kenneth M Anderson. Crisis informaticsnew data for extraordinary times. *Science*, 353(6296):224–225, 2016.
- [93] Kathleen M Carley, Momin Malik, Peter M Landwehr, Jürgen Pfeffer, and Michael Kowalchuck. Crowd sourcing disaster management: The complex nature of twitter usage in padang indonesia. *Safety science*, 90:48–61, 2016.

- [94] Axel Bruns and Yuxian Eugene Liang. Tools and methods for capturing twitter data during natural disasters. *First Monday*, 17(4), 2012.
- [95] Billy Haworth and Eleanor Bruce. A review of volunteered geographic information for disaster management. *Geography Compass*, 9(5):237–250, 2015.
- [96] Flávio EA Horita, João Porto de Albuquerque, Victor Marchezini, and Eduardo M Mendiondo. Bridging the gap between decision-making and emerging big data sources: An application of a model-based framework to disaster management in brazil. *Decision Support Systems*, 97:12–22, 2017.
- [97] Xuan Song, Quanshi Zhang, Yoshihide Sekimoto, Ryosuke Shibasaki, Nicholas Jing Yuan, and Xing Xie. Prediction and simulation of human mobility following natural disasters. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 8(2):29, 2017.
- [98] Yusuke Hara and Masao Kuwahara. Traffic monitoring immediately after a major natural disaster as revealed by probe data—a case in ishinomaki after the great east japan earthquake. *Transportation research part A: policy and practice*, 75:1–15, 2015.
- [99] Qi Want and John E. Taylor. Quantifying, comparing human mobility perturbation during hurricane sandy, typhoon wipha, typhoon haiyan. *Procedia Economics and Finance*, 18:33–38, 2014.
- [100] Melinda Laituri and Kris Kodrich. On line disaster response community: People as sensors of high magnitude disasters using internet gis. *Sensors*, 8(5):3037–3055, 2008.
- [101] Kate Starbird. *Crowdwork, crisis and convergence: How the connected crowd organizes information during mass disruption events*. PhD thesis, University of Colorado at Boulder, 2012.
- [102] Fujio Toriumi, Takeshi Sakaki, Kosuke Shinoda, Kazuhiro Kazama, Satoshi Kurihara, and Itsuki Noda. Information sharing on twitter during the 2011 catastrophic earthquake. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 1025–1028. ACM, 2013.
- [103] Xin Lu and Christa Brelsford. Network structure and community evolution on twitter: human behavior change in response to the 2011 japanese earthquake and tsunami. *Scientific reports*, 4:6773, 2014.
- [104] Lingzi Hong, Myeong Lee, Afra Mashhadi, and Vanessa Frias-Martinez. Towards understanding communication behavior changes during floods using cell phone data. In *International Conference on Social Informatics*, pages 97–107. Springer, 2018.

- [105] Amanda L Hughes, Lise AA St Denis, Leysia Palen, and Kenneth M Anderson. Online public communications by police & fire services during the 2012 hurricane sandy. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1505–1514. ACM, 2014.
- [106] Mats Eriksson and Eva-Karin Olsson. Facebook and twitter in crisis communication: A comparative study of crisis communication professionals and citizens. *Journal of Contingencies and Crisis Management*, 24(4):198–208, 2016.
- [107] Junghoon Chae, Dennis Thom, Yun Jang, SungYe Kim, Thomas Ertl, and David S Ebert. Public behavior response analysis in disaster events utilizing visual analytics of microblog data. *Computers & Graphics*, 38:51–60, 2014.
- [108] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. Crisislex: A lexicon for collecting and filtering microblogged communications in crises. In *ICWSM*, 2014.
- [109] Kate Starbird and Jeannie Stamberger. Tweak the tweet. In *7th International Conference on Information Systems for Crisis Response and Management, ISCRAM 2010*. Information Systems for Crisis Response and Management, ISCRAM, 2010.
- [110] Yury Kryvasheyev, Haohui Chen, Nick Obradovich, Esteban Moro, Pascal Van Hentenryck, James Fowler, and Manuel Cebrian. Rapid assessment of disaster damage using social media activity. *Science advances*, 2(3):e1500779, 2016.
- [111] Andrea L Kavanaugh, Edward A Fox, Steven D Sheetz, Seungwon Yang, Lin Tzy Li, Donald J Shoemaker, Apostol Natsev, and Lexing Xie. Social media use by government: From the routine to the critical. *Government Information Quarterly*, 29(4):480–491, 2012.
- [112] Frances Shaw, Jean Burgess, Kate Crawford, and Axel Bruns. Sharing news, making sense, saying thanks. *Australian Journal of Communication*, 40(1):23, 2013.
- [113] Alexandra Olteanu, Sarah Vieweg, and Carlos Castillo. What to expect when the unexpected happens: Social media communications across crises. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 994–1009. ACM, 2015.
- [114] Kate Starbird and Leysia Palen. *Pass it on?: Retweeting in mass emergency*. International Community on Information Systems for Crisis Response and Management, 2010.
- [115] Adam Acar and Yuya Muraki. Twitter for crisis communication: lessons learned from japan’s tsunami disaster. *International Journal of Web Based Communities*, 7(3):392–402, 2011.

- [116] Patric R Spence, Kenneth A Lachlan, Xialing Lin, and Maria del Greco. Variability in twitter content across the stages of a natural disaster: Implications for crisis communication. *Communication Quarterly*, 63(2):171–186, 2015.
- [117] Jiaying He, Lingzi Hong, Vanessa Frias-Martinez, and Paul Torrens. Uncovering social media reaction pattern to protest events: a spatiotemporal dynamics perspective of ferguson unrest. In *International conference on social informatics*, pages 67–81. Springer, 2015.
- [118] Alan M MacEachren, Anuj Jaiswal, Anthony C Robinson, Scott Pezanowski, Alexander Savelyev, Prasenjit Mitra, Xiao Zhang, and Justine Blanford. Senseplace2: Geotwitter analytics support for situational awareness. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, pages 181–190. IEEE, 2011.
- [119] Zheng Xu, Yunhuai Liu, Neil Yen, Lin Mei, Xiangfeng Luo, Xiao Wei, and Chuanping Hu. Crowdsourcing based description of urban emergency events using social media big data. *IEEE Transactions on Cloud Computing*, 2016.
- [120] George J Borjas, Stephen G Bronars, and Stephen J Trejo. Self-selection and internal migration in the united states. *Journal of urban Economics*, 32(2):159–185, 1992.
- [121] Ronald Skeldon. International migration, internal migration, mobility and urbanization:towards more integrated approaches. *Population Division, Department of Economic and Social Affairs, United Nations*, 2017.
- [122] Douglas T Gurak and Fee Caces. Migration networks and the shaping of migration systems. *International migration systems: A global approach*, pages 150–176, 1992.
- [123] LA Brown and J Holmes. Intra-urban migrant lifelines: a spatial view. *Demography*, 8(1):103–122, 1971.
- [124] Ben CH Kuo. Coping, acculturation, and psychological adaptation among migrants: a theoretical and empirical review and synthesis of the literature. *Health Psychology and Behavioral Medicine: An Open Access Journal*, 2(1):16–33, 2014.
- [125] Yanwei Lin, Qi Zhang, Wen Chen, Jingrong Shi, Siqu Han, Xiaolei Song, Yong Xu, and Li Ling. Association between social integration and health among internal migrants in zhongshan, china. *PloS one*, 11(2):e0148397, 2016.
- [126] Cathy Zimmerman, Ligia Kiss, and Mazeda Hossain. Migration and health: a framework for 21st century policy-making. *PLOS medicine*, 8(5):e1001034, 2011.

- [127] A Bonfiglio. New approaches for researching the determinants of migration processes: Esf strategic workshop on migration research. *International Migration Institute, University of Oxford, for European Science Foundation (ESF)*, 2012.
- [128] Julián Candia, Marta C González, Pu Wang, Timothy Schoenharl, Greg Madey, and Albert-László Barabási. Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical*, 41(22):224015, 2008.
- [129] Dashun Wang, Dino Pedreschi, Chaoming Song, Fosca Giannotti, and Albert-Laszlo Barabasi. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1100–1108. ACM, 2011.
- [130] Emilio Zagheni and Ingmar Weber. You are where you e-mail: using e-mail data to estimate international migration rates. In *Proceedings of the 4th Annual ACM Web Science Conference*, pages 348–351. ACM, 2012.
- [131] Bartosz Hawelka, Izabela Sitko, Euro Beinat, Stanislav Sobolevsky, Pavlos Kazakopoulos, and Carlo Ratti. Geo-located twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, 41(3):260–271, 2014.
- [132] Ingmar Weber, Emilio Zagheni, et al. Studying inter-national mobility through ip geolocation. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 265–274. ACM, 2013.
- [133] Zeynep Tufekci. Big questions for social media big data: Representativeness, validity and other methodological pitfalls. *ICWSM*, 14:505–514, 2014.
- [134] Joshua E Blumenstock. Inferring patterns of internal migration from mobile phone call records: evidence from rwanda. *Information Technology for Development*, 18(2):107–125, 2012.
- [135] Martin T Cadwallader. *Migration and residential mobility: Macro and micro approaches*. Univ of Wisconsin Press, 1992.
- [136] Bruce Frayne and Wade Pendleton. Migration in namibia: Combining macro and micro approaches to research design and analysis. *International Migration Review*, 35(4):1054–1085, 2001.
- [137] Michael Todaro. Internal migration in developing countries: a survey. In *Population and economic change in developing countries*, pages 361–402. University of Chicago Press, 1980.
- [138] Raven Molloy, Christopher L Smith, and Abigail Wozniak. Internal migration in the united states. *Journal of Economic perspectives*, 25(3):173–96, 2011.

- [139] Waldo Tobler. Migration: Ravenstein, thornthwaite, and beyond. *Urban Geography*, 16(4):327–343, 1995.
- [140] S.A. Stouffer. Intervening opportunities: a theory relating mobility and distance. *American sociological review*, 5(6):845–867, 1940.
- [141] David A Plane. Migration space: Doubly constrained gravity model mapping of relative interstate separation. *Annals of the Association of American Geographers*, 74(2):244–256, 1984.
- [142] Shengxiao Li and Pengjun Zhao. Restrained mobility in a high-accessible and migrant-rich area in downtown beijing. *European transport research review*, 10(1):4, 2018.
- [143] June Marie Nogle. Internal migration for recent immigrants to canada. *International Migration Review*, pages 31–48, 1994.
- [144] Martijn Hendriks, Kai Ludwigs, and Ruut Veenhoven. Why are locals happier than internal migrants? the role of daily life. *Social Indicators Research*, 125(2):481–508, 2016.
- [145] Wen H Kuo and Yung-Mei Tsai. Social networking, hardiness and immigrant’s mental health. *Journal of health and social behavior*, pages 133–149, 1986.
- [146] Arul Chib, Holley A Wilkin, and Sri Ranjini Mei Hua. International migrant workers use of mobile phones to seek social support in singapore. *Information Technologies & International Development*, 9(4):pp–19, 2013.
- [147] John W Berry. Immigration, acculturation, and adaptation. *Applied psychology*, 46(1):5–34, 1997.
- [148] Jin Mou, Jinquan Cheng, Sian M Griffiths, Samuel Wong, Sheila Hillier, and Dan Zhang. Internal migration and depressive symptoms among migrant factory workers in shenzhen, china. *Journal of Community Psychology*, 39(2):212–230, 2011.
- [149] Liangliang Cao and Li Fei-Fei. Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.
- [150] Keith Henderson and Tina Eliassi-Rad. Applying latent dirichlet allocation to group discovery in large graphs. In *Proceedings of the 2009 ACM symposium on Applied Computing*, pages 1456–1461. ACM, 2009.
- [151] Joshua Blumenstock, Gabriel Cadamuro, and Robert On. Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264):1073–1076, 2015.

- [152] Mark EJ Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.
- [153] Vanessa Frias-Martinez and Jesus Virseda. On the relationship between socio-economic factors and cell phone usage. In *Proceedings of the fifth international conference on information and communication technologies and development*, pages 76–84. ACM, 2012.
- [154] David M Blei and Michael I Jordan. Modeling annotated data. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 127–134. ACM, 2003.
- [155] Clayton Wukich and Ines Mergel. Closing the citizen-government communication gap: Content, audience, and network analysis of government tweets. *Journal of Homeland Security and Emergency Management*, 12(3):707–735, 2015.
- [156] Sarah Vieweg, Amanda L Hughes, Kate Starbird, and Leysia Palen. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1079–1088. ACM, 2010.
- [157] Starr Roxanne Hiltz, Jane A Kushma, and Linda Plotnick. Use of social media by us public sector emergency managers: Barriers and wish lists. In *ISCRAM*, 2014.
- [158] Teun Terpstra, A De Vries, R Stronkman, and GL Paradies. *Towards a real-time Twitter analysis during crises for operational crisis management*. Simon Fraser University Burnaby, 2012.
- [159] Gennady Andrienko, Natalia Andrienko, Harald Bosch, Thomas Ertl, Georg Fuchs, Piotr Jankowski, and Dennis Thom. Thematic patterns in georeferenced tweets through space-time visual analytics. *Computing in Science & Engineering*, 15(3):72–82, 2013.
- [160] Amittai Axelrod, Xiaodong He, and Jianfeng Gao. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the conference on empirical methods in natural language processing*, pages 355–362. Association for Computational Linguistics, 2011.
- [161] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108. Association for Computational Linguistics, 2010.
- [162] Enhong Chen, Yanggang Lin, Hui Xiong, Qiming Luo, and Haiping Ma. Exploiting probabilistic topic models to improve text categorization under class imbalance. *Information Processing & Management*, 47(2):202–214, 2011.

- [163] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.
- [164] Martin Kulldorff, Richard Heffernan, Jessica Hartman, Renato Assunção, and Farzad Mostashari. A space–time permutation scan statistic for disease outbreak detection. *PLoS medicine*, 2(3):e59, 2005.
- [165] CensusBureau. Usa county shapefiles. <https://www.census.gov/cgi-bin/geo/shapefiles/index.php>, 2016.
- [166] Momin M Malik, Hemank Lamba, Constantine Nakos, and Jurgen Pfeffer. Population bias in geotagged tweets. *People*, 1(3,759.710):3–759, 2015.
- [167] G. Hugo. What we know about circular migration and enhanced mobility. *Migration Policy Institute*, 7, 2013.
- [168] David Akeju. Africa, internal migration. *The Encyclopedia of Human Migration*, 2013.
- [169] Ronald Skeldon. *Migration and development: A global perspective*. Routledge, 2014.
- [170] Lingzi Hong, Enrique Frias-Martinez, and Vanessa Frias-Martinez. Topic models to infer socio-economic maps. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.
- [171] C Hughes, E Zagheni, GJ Abel, A Wisniowski, A Sorichetta, I Weber, and AJ Tatem. Inferring migrations: Traditional methods and new approaches based on mobile phone, social media, and other big data. Technical report, Mimeo prepared for European Commission, 2016.
- [172] G.F. De Jong and J.T. Fawcett. Motivations for migration: an assessment and a value-expectancy research model. *Pergamon Policy Studies on International Development*, 1981.
- [173] Mexican Statistical Institute INEGI. National survey of demographic dynamics 2014. <http://www.beta.inegi.org.mx/proyectos/enchogares/especiales/enadid/2014/default.html>, 2015. Accessed: 2017-06-15.
- [174] Richard D Bedford et al. *New Hebridean Mobility: a study of circular migration*. Canberra, ACT: Dept. of Human Geography, Research School of Pacific Studies, The Australian National University., 2017.
- [175] Raul Montoliu and Daniel Gatica-Perez. Discovering human places of interest from multimodal mobile phone data. In *Proceedings of the 9th International Conference on Mobile and Ubiquitous Multimedia*, page 12. ACM, 2010.
- [176] R. Ahas, S. Silm, O. Järv, E. Saluveer, and M. Tiru. Using mobile positioning data to model locations meaningful to users of mobile phones. *Journal of urban technology*, 17(1):3–27, 2010.

- [177] S. Isaacman, R. Becker, R. Caceres, S. Kobourov, J. Martonosi, M. and Rowland, and A. Varshavsky. Identifying important places in peoples lives from cellular network data. In *Int. Conf. on Pervasive Computing*, pages 133–151, 2011.
- [178] V. Frias-Martinez, J. Virseda, A. Rubio, and E. Frias-Martinez. Towards large scale technology impact analyses: Automatic residential localization from mobile phone-call data. In *4th ACM/IEEE Intl. Conf. Inf. and Communication technologies and development*, page 11. ACM, 2010.
- [179] G.K. Zipf. On the intercity movement of persons. *American sociological review*, 11(6):677–686, 1946.
- [180] F. Simini, M.C. González, A. Maritan, and A.lbert-László Barabási. A universal model for mobility and migration patterns. *Nature*, 484(7392):96, 2012.
- [181] Maxime Lenormand, Aleix Bassolas, and José J Ramasco. Systematic comparison of trip distribution laws and models. *Journal of Transport Geography*, 51:158–169, 2016.
- [182] M. Lenormand, S. Huet, F. Gargiulo, and G. Deffuant. A universal model of commuting networks. *PloS one*, 7(10):e45985, 2012.
- [183] Mexican Statistical Institute INEGI. 2010 census of population and housing units. <http://en.www.inegi.org.mx/proyectos/ccpv/2010/>, 2013. Accessed: 2017-09-12.
- [184] Maxime Lenormand, Miguel Picornell, Oliva G Cantú-Ros, Thomas Louail, Ricardo Herranz, Marc Barthelemy, Enrique Frías-Martínez, Maxi San Miguel, and José J Ramasco. Comparing and modelling land use organization in cities. *Royal Society open science*, 2(12):150449, 2015.
- [185] Yingxiang Yang, Carlos Herrera, Nathan Eagle, and Marta C González. Limits of predictability in commuting flows in the absence of data for calibration. *Scientific reports*, 4:5662, 2014.
- [186] Eunjoon Cho, Seth A Myers, and Jure Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1082–1090. ACM, 2011.
- [187] Nicola Green. On the move: Technology, mobility, and the mediation of social time and space. *The information society*, 18(4):281–292, 2002.
- [188] E. Frias-Martinez and V. Karamcheti. A customizable behavior model for temporal prediction of web user sequences. In *Int. Workshop on Mining Web Data for Discovering Usage Patterns and Profiles*, pages 66–85, 2002.

- [189] Francesco Calabrese, Giusy Di Lorenzo, Liang Liu, and Carlo Ratti. Estimating origin-destination flows using opportunistically collected mobile phone location data from one million users in boston metropolitan area. *IEEE Pervasive Computing*, 10(4):36–44, 2011.
- [190] Donald W Zimmerman and Bruno D Zumbo. Rank transformations and the power of the student t test and welch t’test for non-normal populations with unequal variances. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 47(3):523, 1993.
- [191] Jacob Cohen. A power primer. *Psychological bulletin*, 112(1):155, 1992.
- [192] RM Sakia. The box-cox transformation technique: a review. *The statistician*, pages 169–178, 1992.
- [193] Cati Coe. What is the impact of transnational migration on family life? women’s comparisons of internal and international migration in a small town in ghana. *American Ethnologist*, 38(1):148–163, 2011.
- [194] Jennifer Mason. Managing kinship over long distances: the significance of the visit. *Social Policy and Society*, 3(4):421–429, 2004.
- [195] Dipak Mazumdar. The urban informal sector. *World development*, 4(8):655–679, 1976.
- [196] Jeffrey J Axisa, K Bruce Newbold, and Darren M Scott. Migration, urban growth and commuting distance in toronto’s commuter shed. *Area*, 44(3):344–355, 2012.
- [197] John O Browder, James R Bohland, and Joseph L Scarpaci. Patterns of development on the metropolitan fringe: Urban fringe expansion in bangkok, jakarta, and santiago. *Journal of the American Planning Association*, 61(3):310–327, 1995.
- [198] Steven Vertovec. Cheap calls: the social glue of migrant transnationalism. *Global networks*, 4(2):219–224, 2004.
- [199] George AF Seber and Alan J Lee. *Linear regression analysis*, volume 329. John Wiley & Sons, 2012.
- [200] Joseph F Hair, William C Black, Barry J Babin, Rolph E Anderson, Ronald L Tatham, et al. *Multivariate data analysis (Vol. 6)*. Upper Saddle River, NJ: Pearson Prentice Hall, 2006.
- [201] Prue Holmes and Annick Janson. Migrants’ communication practices with icts: tools for facilitating migration and adaptation? *International journal of technology, knowledge & society.*, 4(6):51–62, 2008.

- [202] GSMA. Gsma country overview: Mexico, mobile driving growth, innovation and opportunity. <https://www.gsma.com/latinamerica/wp-content/uploads/2016/06/report-mexico2016-EN.pdf>. Accessed: 04-10-2018.
- [203] Yves-Alexandre De Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 3:1376, 2013.
- [204] Paul T Jaeger. Disability, human rights, and social justice: The ongoing struggle for online accessibility and equality. *First Monday*, 20(9), 2015.
- [205] Lingzi Hong, Jiahui Wu, Enrique Frias-Martinez, Andrés Villarreal, and Vanessa Frias-Martinez. Accuracy and bias in the identification of internal migrants using cell phone data. In *in the 12th International Conference on Web and Social Media (ICWSM) Workshop on Making Sense of Online Data for Population Research*. AAAI, 2018.
- [206] Daniele Quercia, Michal Kosinski, David Stillwell, and Jon Crowcroft. Our twitter profiles, our selves: Predicting personality with twitter. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, pages 180–185. IEEE, 2011.
- [207] Andrej Zwitter. Big data ethics. *Big Data & Society*, 1(2):1–6, 2014.
- [208] Jalal Mahmud, Jeffrey Nichols, and Clemens Drews. Home location identification of twitter users. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(3):47, 2014.
- [209] Lingzi Hong, Jiahui Wu, Enrique Frias-Martinez, Andrés Villarreal, and Vanessa Frias-Martinez. Characterization of internal migrant behavior in the immediate post-migration period using cell phone traces. In *Proceedings of the Tenth International Conference on Information and Communication Technologies and Development*, page 4. ACM, 2019.
- [210] Randall Wald, Taghi M Khoshgoftaar, Amri Napolitano, and Chris Sumner. Using twitter content to predict psychopathy. In *2012 11th International Conference on Machine Learning and Applications*, volume 2, pages 394–401. IEEE, 2012.
- [211] Munmun De Choudhury, Scott Counts, and Eric Horvitz. Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 3267–3276. ACM, 2013.
- [212] Sean D Young. Behavioral insights on big data: using social media for predicting biomedical outcomes. *Trends in microbiology*, 22(11):601–602, 2014.

- [213] Henry S Farber. Mobility and stability: The dynamics of job change in labor markets. *Handbook of labor economics*, 3:2439–2483, 1999.
- [214] Long Nguyen, Zhou Yang, Jia Li, Guofeng Cao, and Fang Jin. Forecasting people’s needs in hurricane events from social network. *arXiv preprint arXiv:1811.04577*, 2018.
- [215] Robert D Crutchfield, Michael R Geerken, and Walter R Gove. Crime rate and social integration the impact of metropolitan mobility. *Criminology*, 20(3-4):467–478, 1982.
- [216] Xuan Song, Quanshi Zhang, Yoshihide Sekimoto, and Ryosuke Shibasaki. Prediction of human emergency behavior and their mobility following large-scale disaster. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 5–14. ACM, 2014.
- [217] Zhijun Yin, Liangliang Cao, Jiawei Han, Chengxiang Zhai, and Thomas Huang. Geographical topic discovery and comparison. In *Proceedings of the 20th international conference on World wide web*, pages 247–256. ACM, 2011.