POUREBRAHIM, NASTARAN, Ph.D. Human Dynamics in the Age of Big Data: A Theory-Data-Driven Approach. (2019)
Directed by Dr. Selima Sultana. 122 pp.

The revolution of information and communication technology (ICT) in the past two decades have transformed the world and people's lives with the ways that knowledge is produced. With the advancements in location-aware technologies, a large volume of data so-called "big data" is now available through various sources to explore the world. This dissertation examines the potential use of such data in understanding human dynamics by focusing on both theory- and data-driven approaches. Specifically, human dynamics represented by communication and activities is linked to geographic concepts of space and place through social media data to set a research platform for effective use of social media as an information system. Three case studies covering these conceptual linkages are presented to (1) identify communication patterns on social media; (2) identify spatial patterns of activities in urban areas and detect events; and (3) explore urban mobility patterns. The first case study examines the use of and communication dynamics on Twitter during Hurricane Sandy utilizing survey and data analytics techniques. Twitter was identified as a valuable source of disaster-related information. Additionally, the results shed lights on the most significant information that can be derived from Twitter during disasters and the need for establishing bi-directional communications during such events to achieve an effective communication. The second case study examines the potential of Twitter in identifying activities and events and exploring movements during Hurricane Sandy utilizing both time-geographic information and qualitative social media text data. The study provides insights for enhancing situational awareness during natural disasters. The third case study examines the potential of Twitter in modeling commuting trip distribution in New York City. By integrating both traditional and social media data and utilizing machine learning techniques, the study identified Twitter as a valuable source for transportation modeling. Despite the limitations of social media such as the accuracy issue, there is tremendous opportunity for geographers to enrich their understanding of human dynamics in the world. However, we will need new research frameworks, which integrate geographic concepts with information systems theories to theorize the process. Furthermore, integrating various data sources is the key to future research and will need new computational approaches. Addressing these computational challenges, therefore, will be a crucial step to extend the frontier of big data knowledge from a geographic perspective.

KEYWORDS: Big data, social media, Twitter, human dynamics, VGI, natural disasters, Hurricane Sandy, transportation modeling, machine learning, situational awareness, NYC, GIS

HUMAN DYNAMICS IN THE AGE OF BIG DATA:
A THEORY-DATA-DRIVEN APPROACH

by

Nastaran Pourebrahim

A Dissertation Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Greensboro
2019

Approved by

Selima Sultana
Committee Chair

To my late father, Hassan Pourebrahim,

without whom I could not have come this far.

APPROVAL PAGE

This dissertation, written by Nastaran Pourebrahim, has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Chair Selima Sultana

Committee Members Rick Bunch

Paul Knapp

Ming-Hsiang Tsou

____
Date of Acceptance by Committee

____
Date of Final Oral Examination

ACKNOWLEDGMENTS

Completing this dissertation has been one of the most rewarding and challenging experiences of my life. This journey would not have been possible without the support of my mentors, colleagues, friends, and family members. My sincerest gratitude goes to my advisor, Dr. Selima Sultana, for her warm support and constant encouragement from the day I arrived in the U.S. Her words of wisdom on navigating graduate school, patient guidance, and invaluable advice in the past four years made me to accomplish my academic goal. Her intellectual inspiration and devotion to research and teaching have helped me to appreciate science and will have lifetime influence on my academic career.

I would like to thank my committee members, Dr. Rick Bunch, Dr. Paul Knapp, and Dr. Ming-Hsiang Tsou for their continuous guidance and support throughout this process. This dissertation would not have been possible without their input and feedback. I would also like to thank Dr. Zhi-Jun Liu for serving on my proposal committee. My thanks are extended to Dr. Jean-Claude Thill and Dr. Somya Mohanty who shared the data for this dissertation and made valuable suggestions that greatly assisted my research.

I am grateful to all professors at the Department of Geography, Environment, and Sustainability (GES) whose classes gave me lots of benefits and enjoyment. Dr. Marry Hall-Brown deserves a special mention because of her unique character and her devotion to teaching that set a life-long standard for me to follow. I wish to express my gratitude to Lois Carney and James Nelson for facilitating the completion of this dissertation and degree requirements. Lois and James are critical to the success of GES and its graduate students. I would also like to thank all my fellow Ph.D. students at GES whose support made this journey easier. Thank you Joyce, Michele, Nichole, Jennifer, Pankaj, Firoozeh, and Mehrnaz, for your thoughts, well-wishes, e-mails, texts, and advice for my dissertation defense.

I must also thank all my friends in the U.S. who supported me along the way. Many thanks to Mohammad Daneshvar Kakhki who was always helpful in numerous ways during the past four years. Special thanks to Maryam Naghdi, Siavash Pourfazli, Mani Pourfazli, Hamed Bostan, Faraneh Afzal, Milad Moghassem Hamidi, the Rupe family, and most importantly, the Jadidi family who have done a lot to keep me going.

I would like to express my deepest gratitude to my family, especially my mother, Nosrat Khajeh, for loving and supporting me to no end and showing me the true meaning of the "death of distance". I am extremely grateful to my sisters, Narges and Nasim, my brother-in-law, Arash, and my nephew, Ilya, for always being by my side, in moments of joy and sadness. Thanks should also go to my in-laws, Forough Kazemi, Elham Amirazar, and Arash Amirazar, for their support. Lastly but most importantly, I would like to thank my husband, Armin, for taking this long journey with me, encouraging me to keep dreaming, and being a true friend when I struggled. No words can describe my love and thanks to him.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

CHAPTER I

INTRODUCTION

1.1 Research Background and Motivations

The emergence of information and communication technology (ICT) and advancements in location-aware technologies in the past two decades have radically changed the way people live and communicate (Shaw, Tsou, & Ye, 2016). The growth of Geo Web and geographic information contributed by users through different interfaces has created a new role for Geographic Information Systems (GIS) as social media (Sui & Goodchild, 2011). With the smart phones and mobile devices at fingertips, people leave their digital footprints on the cyberspace (i.e., social media, web pages) and produce a high volume of data so-called “big data” (Tsou, 2015). Much of these big data provide a rich body of geographic information in forms of latitude and longitude named “geotags” (Yang, 2015). Since these data are voluntary and public, they can be used to support scientific inquiries in various disciplines, including geography (Yang, 2015).

The research on social media has been primarily focused on content and social network analysis or event detection in various contexts (i.e., emergency, health, and traffic management) by using semantic and spatio-temporal information of the data (Steiger, de Albuquerque, & Zipf, 2015). Recently, there has been more research on developing computational methods to capture and infer relevant information from social media (i.e., Huang, Li, Wang, & Ning, 2019; Mazhar Rathore, Ahmad, Paul, Hong, & Seo, 2017). Nevertheless, the prime criticism of research with big data has been their failure in adherence to theoretical frameworks (Shelton, Poorthuis, Graham, & Zook, 2014). It has been simply assumed that these new datasets could generate more insightful results with complicated algorithms than the conventional theoretical approaches (Shelton et al., 2014)—a notion known as “the end of theory” (Anderson, 2008).

These data-driven approaches have also been criticized for ignoring the need for any domain expertise in the analyzed subjects (Shelton et al., 2014). Miller (2010), however, argues that these approaches can benefit from the domain expertise and theory to discover patterns in the data that otherwise would not have been seen. Geographers and GIS scientists are among the most prominent experts that can provide new insights from the analysis of such data in the urban context with the help of geographic concepts. With the ever-growing production and use of these datasets, there is a need to explore which data to collect and how to ask and answer questions to exploit valuable patterns and unprecedented insights of these data into the world (Shelton et al., 2014; Floridi, 2012). Hence, the main questions are: what are the potential uses of social media in understanding the world around us? And how could they be used effectively? The answer to these questions, however, is not possible without using existing theories and concepts and developing new conceptual frameworks to analyze such data. This dissertation attempts to fill a gap in previous studies that utilized social media data in urban context, but often avoided conceptualizing the process. The aim of this dissertation, therefore, is to set a platform for research on effective data-driven use of social media utilizing existing information systems (IS) theories and geographic concepts.

1.2 Effective Use of Information Systems in Geographic Context

The aim of Information systems (IS) research has long been the advancement of effective use and effective utilization of information technology by individuals, groups, organizations, society, and nations to improve economic and social welfare (Agarwal, 2012). The shift from system use to effective use, therefore, is the priority of IS research (Marcolin, Compeau, Munro, & Huff, 2000). Burton-Jones & Grange (2013) developed the effective use theory (EUT) to identify the nature and drivers of effective use employing the research on system use and representation theory (RT). According to RT, three elements of use are motivations of users, nature and purpose of systems, and the characteristics of tasks. Users leverage digital representations of the systems to enhance their understanding from real-world domains and take actions to achieve their goal (Bonaretti & Piccoli, 2019). To obtain these representations (the deep structure of IS), users must have access to surface (facilities allowing users' interaction with the representations) and physical structures (machinery supporting other structures) (Burton-Jones & Grange, 2013). The dimensions of effective use, therefore, are transparent interaction, representational fidelity, and informed action. Transparent interaction is the extent to which users can access the system's representation by its surface and physical structure; representation fidelity is the extent to which users obtain faithful representation of the domain; and informed action is the extent to which users act upon these representation to improve their state (Burton-Jones & Grange, 2013).

Burton-Jones & Volkoff (2017) discuss developing context-specific theories of effective use. Their approach applies the concepts of affordance network (to understand the intended outcome or purpose) and affordance actualization (to understand the associated actions and interactions). By focusing on immediate concrete outcomes, the complexity of effective use of systems can be decomposed to manageable pieces. The effective use, therefore, can be conceptualized and contextualized as set of affordances that users actualize for a given system (Figure 1.1). We can understand effective use of an information system in a specific context by identifying different goals of a system, the way the network of affordances links together, and the way to actualize each affordance. Applying these concepts to social media as an information system, the major goal of using social media information system in urban setting from a decision maker perspective could be summarized to: (1) establish an effective communication with the urban residents for sharing information and (2) become aware of what is happening in the urban area for a better urban management, which I call human dynamics awareness.

Humans participate in a large number of activities ranging from electronic communications to engaging in entertainment and work activities (Barabási, 2005). Understanding these activities could benefit a number of stakeholders such as urban planners, transport engineers, health and disaster managers (Candia et al., 2008). Given the major goal of social media information system as effective communication and human dynamics awareness, understanding communications and activities, therefore, are the immediate outcomes that need to be achieved for an effective use of such system (Figure 1.1). To achieve these outcomes, the user (i.e., researcher, analyst) will need to actualize a set of affordances including collecting, processing, and analyzing the data. Although actualizations of these affordances and mapping them are not the aim of this dissertation, I will discuss these dimensions in the conclusion chapter to give an insight for future research.

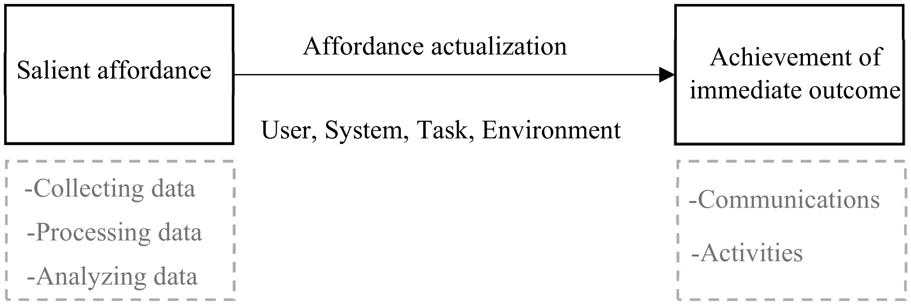


Figure 1. 1. Affordance-Outcome Model of Effective Use (Adopted from Burton-Jones & Volkoff, 2017).

An important difference of effective use in the urban context compared to its original context is the urban environment and how it affects the elements of use (users, systems, and tasks). In the original context of the theory, users, systems, and tasks are generally stable and the represented domain is not reshaped by the environment (Bonaretti & Piccoli, 2019). However, human dynamics shape the environment and are shaped by it. Since effective use is about faithful representation of the real-world domain (human dynamics in urban setting in our case), environment should be added to the framework as an element of use (Figure 1.1).

Environment can be presented as space and place, the two fundamental concepts in geography, and more broadly in social sciences, humanities, and information science (Yang, 2015). The big data produced through social media not only provide a platform for researchers to study people communications in cyberspace, but also establish a dynamic record of people activities in realspace. Geographic information systems (GIS), have been mostly dominated by space, a generic concept represented by Cartesian coordinates (Sui & Goodchild, 2011). Place, however, has less been the focus of GIS. The recent convergence of GIS and social media (Sui & Goodchild, 2011) also indicates a need for theoretical works to reconcile the worlds of space and place (Yang, 2015). Thus, this research addresses the challenge by linking human dynamics with space and place through social media information system (Figure 1.2).

Human dynamics consist of two parts, (1) communications and (2) activities—general physical activities of people in space and their movements. Space in this study refers to both cyberspace and realspace. From a geographic perspective, realspace is about geometrics, spatial dimensions, and the world of objects. Place, however, can be understood and clarified through the meaning given by people (Tuan, 1977) or through movements. It is identified through a “bottom-up” view that captures the local environment, human activities, and experiences (Yang, Ye, & Sui, 2016). Place in this study, therefore, can be absolute such as a specific street, home or work census tracts, and place of an event as experienced by people or it can be relational, where place is identified through its relationship with other places either through people movements or experiences. Figure 1.2 demonstrates the research framework, where human dynamics is linked to space and place through social media information system. These links can be described as:

(1) Human dynamics (communication), Space (cyberspace), Social media information system: To understand communications in cyberspace utilizing social media text and user information, and survey.

(2) Human dynamics (activities), Space (realspace), Social media information system: To understand spatial patterns of human activities in realspace utilizing geotagged information.

(3) Human dynamics (communications and activities), Place, Social media information system: To identify absolute and relational places by integrating communications (link 1) and activities (link 2) utilizing social media text and geotagged information; and to identify relational places generated by human movements utilizing geotagged information.

(4) Human dynamics (activities), Place, Social media information system: To understand activities (movements) generated by absolute and relational places utilizing geotagged information.

The relationship between human dynamics and space is explained in links 1 and 2. Link 1 addresses potential of social media as a means of communication, where source of information, discussions, and users' networks can be explored. Links 2 addresses the spatial patterns of activities in realspace, where cluster of activities in urban areas can be identified through social media. A two-way relationship exists between human dynamics and place, where either human dynamics generate place or place generates human dynamics. Link 3 addresses the generation of place through (1) integrating textual communications on cyberspace (link 1) with the spatial patterns of activities (link 2) and (2) human movements, where places of high/low human flows can be identified. These generated places can be explored from both data and theory perspectives. From a data perspective, these places are the locations of events or high/low human flows in urban areas showing the potential of social media for situational awareness. From a theory perspective, however, these places are created because of people experiencing an event and giving a meaning to it at a specific space and time or because of human movements. Link 4 explores human movements generated by places (i.e., home and work census tract). Since the generation of movements is related to the characteristics of these places, this link demonstrates a modeling approach to explore the potential of social media in understanding movements.

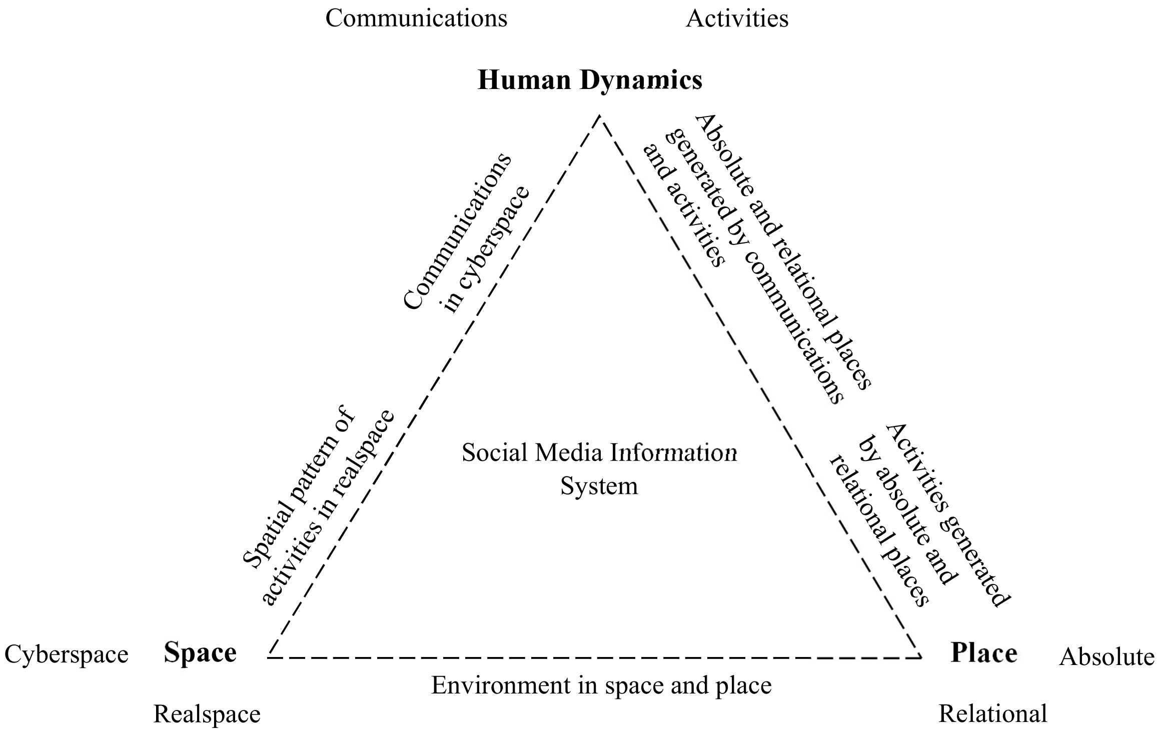

Communications
Activities
Human Dynamics
Communications in cyberspace
Absolute and relational places generated by communications and activities
Spatial pattern of activities in realspace
Social Media Information System
Activities generated by absolute and relational places
Cyberspace
Space
Realspace
Environment in space and place
Place
Absolute
Relational

Figure 1. 2. Framework for Understanding Human Dynamics by Social Media Information System.

1.3 Research Objectives

The ever-expanding use of big data generated through various sources such as social media has provided us with the opportunity to understand human dynamics. While big data have been utilized in different research, theorizing the process from a geographic perspective has less been investigated. The major objective of this dissertation, therefore, is to contribute to existing theories and methods of developing a framework of utilizing social media to understand human dynamics through the lens of geographic concepts. Employing three data-driven case studies to reflect the identified conceptual links, the objectives of this research are to:

(1) Assess potential of social media as a means of communication during events.

(2) Examine potential of social media in identifying activities and events in urban areas.

(3) Examine potential of social media in modeling movements in urban areas.

1.4 Synopsis of Dissertation

The dissertation is organized as follows: Chapter I is an introduction to the study and outlines the research problem and objectives. By reviewing existing theories and concepts, I setup the conceptual framework of this dissertation. Chapter II examines the use of Twitter during Hurricane Sandy by a survey of general population in affected areas. Communication dynamics on Twitter are also explored utilizing tweets generated on entire Twitter platform in general and Hurricane Sandy impacted areas in particular. This chapter addresses link 1 in the conceptual framework. Chapter III examines the potential of Twitter in identifying activities and events and exploring movements in New York City during Hurricane Sandy from both data and theory perspectives. This chapter addresses links 3 and 4 in the conceptual framework. Chapter IV examines the potential of Twitter in modeling commuting trip distribution in New York City. This chapter addresses link 4 in the conceptual framework. Chapter V draws the conclusions, indicates the limitations, and provides potential future research.

CHAPTER II

UNDERSTANDING COMMUNICATION DYNAMICS ON TWITTER DURING NATURAL DISASTERS: A CASE STUDY OF HURRICANE SANDY [1]

2.1 Introduction

Climate change is a major threat of our time and is expected to intensify the frequency of extreme weather events (Nam, Hayes, Svoboda, Tadesse, & Wilhite, 2015). Providing emergency information to population living in the vulnerable areas, therefore, has become a measure of disaster resilience and a policy priority(Feldman et al., 2016). With nearly 2 billion Facebook active users, 6 billion YouTube videos viewed and approximately 700 million tweets posted on Twitter every day, social media platforms have become the major channels for people to communicate and stay informed (Internet Live Stats; Mohammadi, Wang, & Taylor, 2016). Given the increasing presence of social media in everyday life, it can be a major platform for sharing emergency information such as warnings, disaster relief efforts, crisis mapping for escape routes, search and rescue, and connecting community members following a disaster (Kim & Hastak, 2018; Qadir et al., 2016; Houston et al., 2015).

Twitter is one of the prime social media platforms, which has been used not only during emergency situations, but it also changed the way people create, disseminate, and share emergency information (Lifang Li, Zhang, Tian, & Wang, 2018). The real-time characteristic of Twitter makes it a suitable crowdsourcing platform for dissemination and collection of

[1] Pourebrahim, N., Sultana, S., Edwards, J., Gochanour, A., & Mohanty, S. (2019). Understanding communication dynamics on Twitter during natural disasters: A case study of Hurricane Sandy. *International Journal of Disaster Risk Reduction*, *37*(May), 101176. https://doi.org/10.1016/j.ijdrr.2019.101176

information including texts and pictures during disasters and crisis events (Yuan & Liu, 2018), which enhances the public awareness of a situation instantly. One of the major challenging issues facing emergency officials is the development of warning methods for residents at risk in order for them to take appropriate actions immediately (Sorensen, Sorensen, Smith, & Williams, 2009). Twitter use has been growing rapidly especially during disasters by the local officials (Tyshchuk, Hui, Grabowski, & Wallace, 2012), but the effectiveness of this system is an on-going debate. Hence, it is important to understand its potential as a mean of communication during disaster since it informs people and in turn might influence their actions in preparing for a natural disaster (Feldman et al., 2016). While the potential role of Twitter during disaster has been discussed, the effectiveness of widespread use of Twitter for receiving and disseminating risk information during such events has been less investigated in academic research (Ragini, Anand, & Bhaskar, 2018; Feldman et al., 2016;). Without evidence-based strategies of how these technologies are being used, the implementation of social media tools in disaster management remains challenging (Feldman et al., 2016).

Identifying the appropriate ways to use and analyze social media data is an important task for researchers to draw reliable conclusions so that the full potential of social media during emergencies can be achieved (Lindsay, 2011). The importance of content analysis and social network structure during disaster has been identified in the literature (Jooho Kim, Bae, & Hastak, 2018), but application of Twitter data in academic research is still at its infancy. Given the anticipated increase in frequencies of natural hazards such as hurricanes and their resulting damages in the United States in the coming decades (Eshghi & Schmidtke, 2018), better understanding of social media data is crucial so that it can be utilized by emergency authorities. Examining the structures of the online social networks, the underlying mechanism of online users' behaviors and their shared contents, therefore, is the key to achieve this goal (Jooho Kim et al., 2018).

This study investigates the usage of Twitter during Hurricane Sandy by conducting a general population survey of the affected regions and analyzing the content/social network structure of messages and users on Twitter during this period. Our objectives are to (1) Identify the sources of information received by and shared with the coastal areas' residents affected by Hurricane Sandy, (2) examine the Twitter users' involvement in discussions about Hurricane Sandy and the content of messages shared before, during, and after the hurricane, and (3) examine the social network structure of Twitter users before, during, and after the hurricane. The paper is organized as follows: we first present a review of relevant literature regarding social media data and disaster management in section 2.2 followed by discussions on study site in section 2.3, and data collection and methods in section 2.4. We then discuss the results of our analysis in section 2.5. We present conclusions, limitations and suggestions for future research in section 2.6.

2.2 Social Media Analysis in Disaster

Social media platforms such as Twitter allow public and officials to share texts and photos, which can be a powerful means of communications during disasters. Sharing and transferring local knowledge in all stages of crisis (pre, during, post) and building social capital are essentials for communities' resiliency (Grube & Storr, 2014; Joshi & Aoki, 2014). Social media can facilitate this process since a number of people, disaster-affected communities, and organizations are linked via these online networks (Jooho Kim et al., 2018). Research shows that local city officials' evaluations of their ability in controlling a crisis and the strength of their responses are positively related to the extent of social media they use (Graham, Avery, & Park, 2015). People also expect fast arriving of help after posting a request on a social media site (B. F. Liu, Fraustino, & Jin, 2015). With the rapid growing of social media as emergency communication channels (Jooho Kim et al., 2018), various scholars have investigated the potential of generated big data using different techniques. For example, the potential use of Twitter during disasters such as floods (Ann St Denis, Palen, & Anderson, 2014; Murthy & Longwell, 2013), earthquakes (Jung & Moro, 2014; Muralidharan, Rasmussen, Patterson, & Shin, 2011), hurricanes (Hughes, St. Denis, Palen, & Anderson, 2014), tornados (Ukkusuri, Zhan, Sadri, & Ye, 2014), tsunamis (Acar & Muraki, 2011), wildfires (Sutton, Palen, & Shklovski, 2008), volcanic hazards (Chatfield & Reddick, 2015), and droughts (Tang, Zhang, Xu, & Vo, 2015) has been investigated.

The shared information on social media platforms such as Twitter are useful for public and emergency management authorities to understand on-the-ground realities during emergencies (Ukkusuri et al., 2014). Information exchange behavior of social media users, their various spatial and temporal activity patterns on social media, and the content of shared messages are examples of such useful information. These behaviors and activities have been reported to change based on the crisis life cycle, affected regions, event's type and characteristics (Lifang Li et al., 2018; Martín, Li, & Cutter, 2017). Social media contents during disasters have been analyzed using text mining and sentiment analysis. Text mining is becoming a popular method to understand the unstructured text information (Jooho Kim et al., 2018). A word cloud is commonly used to determine the most frequently used words and illustrate a visual representation of text data generated in social media (Jooho Kim et al., 2018). However, to understand the true meaning of the text, sentiment analysis has been mainly applied to classify texts into positive and negative (J. Rexiline Ragini et al., 2018). Researchers have become more interested in sentiment analysis during emergency situations using machine learning techniques. Ragini, Rubesh Anand, & Bhaskar (2018), for example, used support vector machine (SVM) to classify the tweets and their positive or negative sentiments of individual's needs during different disasters. In another study, Vo & Collier (2013) identified various emotions such as anger, calm, unpleasantness, sadness, anxiety, fear and relief during four of the Japan's earthquakes. The content analysis of social media data coupled with spatial and visual analysis techniques are becoming a major research area to understand potential of such data in disaster management (Tsou et al., 2017).

Another line of study related to social media is exploring online users' behaviors and interactions through social network analysis. Social network analysis is used to study network structures, relationship properties of networks, communication's patterns between users, the role of various actors in the network, and community detection (Williams, McMurray, Kurz, & Hugo Lambert, 2015; Park, Lim, & Park, 2015). Graph theory is the major approach in social network analysis where a network of nodes and links is created to examine the relationship among social media users (Sapountzi & Psannis, 2018). Three main analyses found in literature are influence analysis, link analysis, and community detection (Sapountzi & Psannis, 2018). Studies have used various metrics (e.g., degree centrality) and algorithms (e.g., modularity) to explore online network structures, communities and the users' engagement during disasters (Jooho Kim et al., 2018; Jooho Kim & Hastak, 2018). As the number and expectations of social media users during disasters continue to grow (B. F. Liu et al., 2015), the sources and types of information received by and shared with people and the functioning of social media during extreme whether events need more investigation particularly from a social network perspective. Data generated by social media such as Twitter has been applied in various fields of academic research such as urban and transportation planning and public health (Pourebrahim, Sultana, Niakanlahiji, & Thill, 2019; Karduni et al., 2017; Allen, Tsou, Aslam, Nagel, & Gawron, 2016). Its application in disaster management research, however, is still at its early stages (Ragini et al., 2018) and this study intends to fill that gap in the disaster management literature.

2.3 Context of the Study

Hurricane Sandy started on October 22, 2012, moved from the Caribbean to the U.S. Eastern Seaboard, and finally made landfall near Brigantine, New Jersey, around 8:00 p.m. on October 29, 2012 (Figure 2.1) (National Weather Service, 2013). Sandy caused 147 direct deaths, and around 650,000 damaged or destroyed houses, and left approximately 8.5 million customers without power during and after the storm (National Weather Service, 2013), making it one of the deadliest and most destructive hurricanes in the history of the United States (Saleem, Xu, & Ruths, 2014a). New York, New Jersey and Connecticut, especially in and around the New York City metropolitan received the record levels of storm surges (Figure 2.2)(National Weather Service, 2013). Despite heavy power outages and disruptions, people used social media such as Twitter as a critical platform to express the impacts of Sandy on their lives and material goods (Stewart & Gail Wilson, 2016; Shelton, poorthuis, Graham, & Zook, 2014; Guskin & Hitlin, 2012) . The hurricane-related tweets that were posted during the Sandy across the entire Twitter platform, and the geo-tagged Twitter data generated in the affected coastal counties in New York, New Jersey and Connecticut are used for our research.

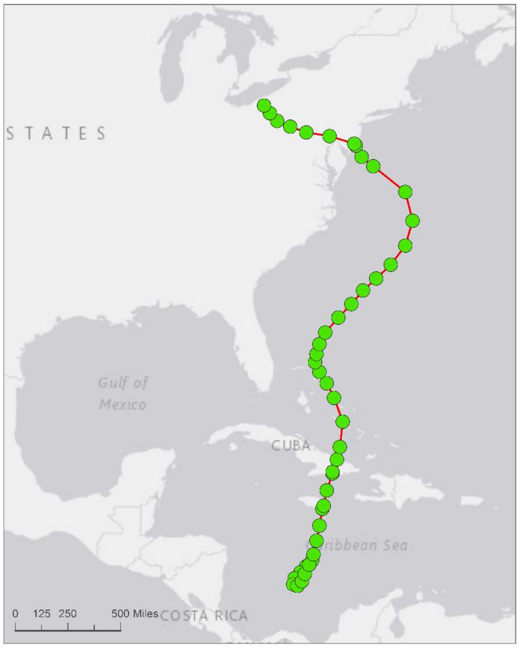


Figure 2. 1. Hurricane Sandy Track. (Source: National Hurricane Center).

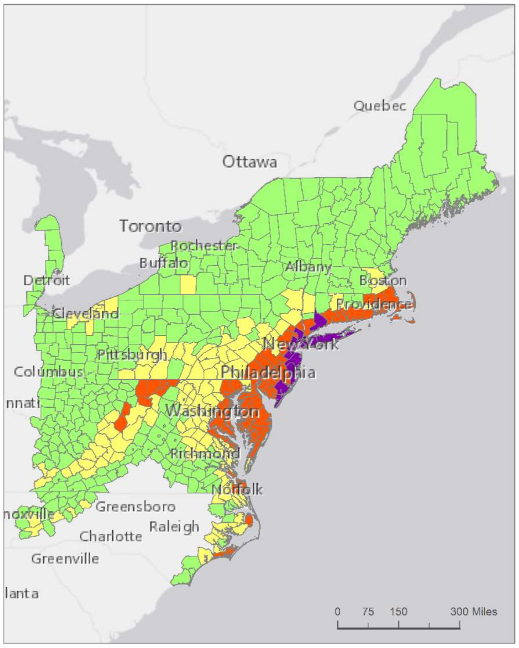


Figure 2. 2. Hurricane Sandy Impact Analysis. (Source: FEMA Modeling Task Force).

2.4 Data and Methods

In addition to Twitter data, this study includes survey data. For the first research objective (section 5.1), a two-part survey of respondents who lived in the coastal counties of Connecticut, New Jersey, and New York at the time of Hurricane Sandy was conducted (Figure 2.3). The first phase of data collection involved the use of a telephone-based survey. Since we were highly interested in the ways respondents used Twitter during Hurricane Sandy, the telephone-based survey was supplemented with a web-based survey of respondents recruited via Twitter. The web-based survey of Twitter users also helped to ensure a large enough sample of Twitter users so that meaningful comparisons between Twitter users and non-Twitter users can be made. The data collection period was started on January 28, 2015 and ended on May 7, 2015.

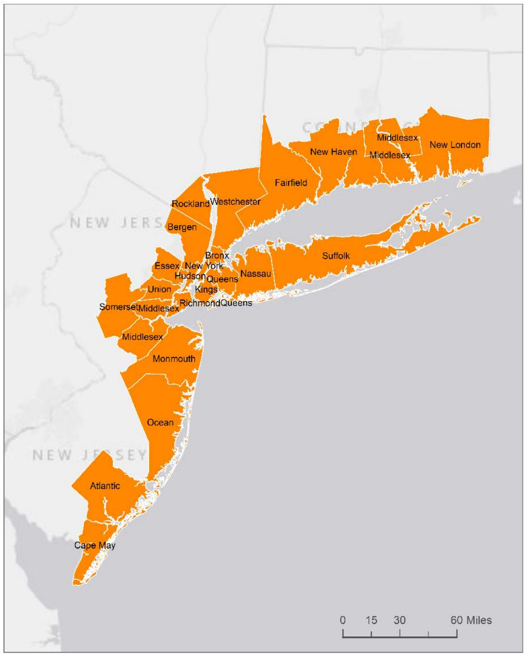


Figure 2. 3. Survey Catchment Area.

2.4.1 Telephone-Based Survey

The first phase of data collection involved a telephone-based survey of a representative sample of residents from 23 counties (Figure 2.3) in Connecticut (Fairfield, Middlesex, New Haven, New London), New Jersey (Atlantic, Bergen, Cape May, Essex, Hudson, Middlesex, Monmouth, Ocean, Somerset, Union), and New York (Bronx, Kings, Nassau, New York, Queens, Richmond, Rockland, Suffolk, Westchester). Respondents were screened to ensure that they were residents of one of the appropriate states and counties at the time Hurricane Sandy made its landfall. The telephone numbers were purchased from Survey Sampling International. A dual-frame (cellphone & landline), random-digit-dialing (RDD) sample was purchased that included a random sample drawn from the universe of existing telephone numbers assigned to the catchment area. The final sample size for those recruited and surveyed via telephone was 514, with a cooperation rate of 12.4%. We think that the relatively low cooperation rate is likely due to the length of time that passed between Hurricane Sandy and the data collection period (approximately 2 ½ years). The low cooperation rate might have also been influenced by respondents' reluctance due to the large number of research projects that have attempted to survey the local population in the years following Hurricane Sandy. Regarding the sample of respondents recruited via telephone, the demographic characteristics (Table 2.1) were evenly split according to gender, with the sample comprised of 49.2% men and 49.8% women. The racial composition of the sample was comprised of 51.6% white, 18.9% black, 5.1% Asian, 12.5% "others" and 3.7% multiracial. The level of education for the respondents was above average compared to the national population, with about half of the sample holding a bachelor's degree or a graduate degree. The survey included questions about the source of information received by residents during the hurricane and the types of information they shared. Example of survey questions are represented in Table 2.2.

Table 2. 1 Demographic Characteristics of Telephone and Web Surveys.

	Gender		**Racial Composition**					**Education**
	Male	Female	White	Black	Asian	Others	multiracial	Bachelor/Graduate degree
Telephone Survey	49.2%	49.8%	51.6%	18.9%	5.1%	12.5%	3.7%	50%
Web Survey	57%	43%	57.3%,	3.9%	6.8%	3.9%	5.3%	50%

Table 2. 2 Example of Survey Questions.

Survey Questions
During Hurricane Sandy, did you get any weather-related information from the following sources (Check all that apply)?
In what ways did you receive weather-related information?
Of the ways you received weather-related information, from which did you get *most* of your information?
During the storm, did the way that you received most of your information change? If yes: a) After it changed, how did you get most of your information? b) Why did the way you received most of your information change during the storm?
At any time during Hurricane Sandy, did your household lose access to:
During Hurricane Sandy, did you tweet or retweet storm-related information? If yes: a) What was the nature of the information that you tweeted or retweeted? b) What was the source of the information that you tweeted or retweeted? c) What form of information did you tweet or retweet?
Do you follow any of these sources of information on Twitter?

2.4.2 Web-Based Survey

In the second phase of data collection, participants were recruited through Twitter. We used all geo-tagged tweets sent from the catchment area (Figure 2.3) before, during, and directly following Hurricane Sandy's landfall to identify Twitter users. From this data, approximately 26,000 unique Twitter users were identified, 20,000 of which were randomly selected for inclusion in the web-based survey sample. A Twitter message was sent as a “@ reply” to each Twitter user. When participants clicked a link embedded in the tweet, they were directed to a web-based survey, which was virtually identical to the telephone survey. The 20,000 tweets generated 207 click-throughs for a total of 170 completed surveys, resulting in a response rate of slightly less than one percent. Much like the telephone sample, this low response rate is likely attributable to the length of time between the event and the data collection period, and to the respondent's reluctance. Furthermore, Twitter-based survey recruitment is a relatively novel approach and users may have been less willing to consider surveying a legitimate use of Twitter's platform. Regarding the sample of respondents recruited via Twitter, the demographic characteristics were as follows (Table 2.1): The modal category for gender was men, with men making up 57.0% of the sample and women comprising 43.0% of the sample. Regarding race, the vast majority of the sample was white 57.3%, followed by 6.8% Asian, 5.3% multi-racial, 3.9% black, and 3.9% “others”. The education level of respondents was above average compared to the national population, with about half of the sample holding a bachelor's degree or a graduate degree.

2.4.3 Twitter Dataset

A total of 13.7 million Twitter messages were collected from Oct. 22 to Nov. 7, 2012 using Firehose streaming API via GNIP (Gnip APIs). The raw data were indexed and inserted into a distributed NoSQL (MongoDB) database for storage. This database served as the central repository of data for all subsequent analyses. We created two datasets based on: (1) Keyword – Twitter messages matching a set of Sandy-related terms comprised of keywords, hashtags, and user names (9.3 million Twitter messages); and (2) Geo-tagged – Twitter messages from New York, New Jersey, and Connecticut (4.4 million Twitter messages). Both datasets (keyword and geo-tagged) were further divided into three temporal phases: (1) Pre-hurricane (10/22/2012–10/28/2012), (2) During-hurricane (10/29/2012–10/31/2012), and (3) Post-hurricane (11/01/2012–11/07/2012).

2.4.4 Temporal Analysis

Identifying temporal trends during a disaster is an important method to determine the tracking system of tweets over different phases of the disaster (Murthy & Gross, 2017; Chatfield, Scholl, & Brajawidagda, 2014;). The collected datasets contained metadata attributes of the time at which the messages were posted on the Twitter network. The time-stamp attribute has a resolution of milliseconds in relation to the GMT time zone. Analysis on the keyword and the geo-tagged datasets were conducted to illustrate the peaks and valleys in the data in order to better understand the involvement of Twitter users in discussions about Hurricane Sandy during the three identified phases. The data were aggregated by number of messages per-hour and the number of unique users' per hour to visualize the results in the temporal zones.

2.4.5 Klout Analysis

Each user captured by the data, contains the metadata attribute of Klout score. Klout score is a metric to measure influence of users on online social networks (Rao, Spasojevic, Li, & Dsouza, 2015). Klout score of a user is measured based on three components including: true reach, which measures how many people a user influences; amplification, which refers to how much the user influences them; and network impact that measures the influence of the user's network (Edwards, Spence, Gentile, Edwards, & Edwards, 2013). This score for a Twitter user is a numerical value from 1 to 100, with a higher score representing a higher level of influence. It is based on the size of user's social network (friends, followers) and correlates with the reactions to the user's posting by other Twitter users. In this research, the scores pertaining to individual users were aggregated by hour to understand the involvement of influential users in the discussions about Hurricane Sandy. The data were then compared across the temporal zones for both the keyword and the geo-tagged datasets. It should be noted that the Klout scores were collected after the Hurricane Sandy and there might have been changes in the scores during that time. The Klout score service provider was shut down later and there is no way to verify the information.

2.4.6 Text Analysis

The contents of the Twitter messages were analyzed using word cloud and word collation/co-existence in order to understand the key topics that were discussed in the temporal phases of the study. The text contents of the tweets for the different phases of the hurricane from both datasets were extracted. The text was then processed for cleanup (removal of stopwords, hyperlinks, common terms), and word stemming was utilized for abbreviations in the short messages. Word importance weights were calculated using Term Frequency – Inverse Document

Frequency – TF-IDF (Salton & Buckley, 1988) for individual words, where each word's weight is an importance metric calculated across the corpus of all documents (in our case tweets). The values were represented in word clouds for both keyword and geo-tagged datasets. Word co-existence analysis was done by examining existence of bigrams of words present in Twitter messages. For all pairs of collated words, a graph was constructed using the word pairs as nodes in the graph with edges/links denoting a co-existence connection. Weights were used for the nodes and edges representing the repeated occurrence of entities and pairs. The resulting network was analyzed and visualized for connectedness (degree), influence (betweenness centrality) and community memberships. Louvain modularity (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008) was used for detection of communities of words. The method utilizes recursive grouping of smaller communities by representing them as nodes and evaluating how dense the connections are while maximizing/optimizing for modularity. The high positive value of modularity measure indicates the presence of community structure with dense connected nodes within the partition sets (Croitoru, Wayant, Crooks, Radzikowski, & Stefanidis, 2015). The analysis of the graph was performed using Gephi, an open source software for network analysis and visualization (Heymann & Grand, 2013).

2.4.7 Sentiment Analysis

Sentiment analysis is an ongoing field of research determining the positive or negative opinion of people about a text message, a specific entity, or a topic in general (Sapountzi & Psannis, 2018). Support vector machine (SVM), a common machine learning method for classification (Karimi, Sultana, Shirzadi Babakan, & Suthaharan, 2019; Ragini et al., 2018; Neppalli, Caragea, Squicciarini, Tapia, & Stehle, 2017) was used to determine polarity of the messages (positive, negative, or neutral) sent during the three phases of Hurricane Sandy. The process includes various techniques including text cleanup (stop-words removal, URL removal, emoticon filtering), feature-extraction (context features, natural language processing, rare word filter, stemming), and vectorization of text. The model used for the analysis of the present datasets is based on a trained dataset of approximately 4.2 million codified tweets in each category and is purpose-built for codifying short text messages such as tweets with an accuracy of 84% (10-fold cross-validation with 2.1 million codified tweets). The model also resulted in 81% accuracy in a double-blind verification by human coders. The sentiment model assigns each tweet a value between −2.0 and +2.0, where −2.0 to −1.0 is codified as negative, −0.99 to +0.99 is codified as neutral and +1 to +2 is codified as positive. Both the keyword and geo-tagged datasets were codified with sentiment values for each tweet and average sentiment values were aggregated by hour.

2.4.8 Social Network Analysis

Social network analysis uses graph theory to understand social phenomenon through the relationship among people, groups, and things (Hansen, Shneiderman, & Smith, 2010). A social network is represented by a graph of nodes and links, where nodes are individual actors and links are social ties, relationships, exchanges, or interactions among actors (Chatfield & Brajawidagda, 2012). The user network analysis was performed based on the metadata of user mentions in Twitter Messages for keyword dataset. Twitter users can mention other users by posting a message using the format of '@username' to reference a particular user or reply to another user's tweet. A directed graph was created for each mention of the user and edges were added from the originating user to the mentioned user. Different network metrics (for each stage of the hurricane) were calculated including average degree, average weighted degree, graph density, network diameter, centrality (in degree, out degree, betweenness, eigenvector), average clustering coefficient, and average path length. A full description of these measures can be found in Jackson (2008). Louvain method (Blondel et al., 2008) in Gephi was used to calculate modularity and divide the graph into groups of users who are well connected or clustered together in the graph known as communities. The clusters were visualized by Force Atlas algorithm (Bastian, Heymann, & Jacomy, 2009).

2.5 Results

2.5.1 Research Objective 1: To Identify the Sources of Information Received by and Shared with the Coastal Areas' Residents Affected by Hurricane Sandy

The telephone and web-based survey instruments were used to understand the use of Twitter as a communication platform during Hurricane Sandy. The following three research questions were addressed from the survey: (1) How did people obtain information during Hurricane Sandy? (2) From whom did people obtain information during Hurricane Sandy and how did these sources differ between Twitter users and non-Twitter users? and (3) What type of information did Twitter users share during Hurricane Sandy?

2.5.1.1 Telephone and web-based survey results

In both the telephone and web-based surveys, individuals were asked whether or not they sought information through ten different mediums. Of particular interest was the difference between Twitter and non-Twitter users in receiving information. Utilizing Chi-Square test, we found statistically significant differences between Twitter users and non-Twitter users on four of the mediums including text message with X2 (1, N=667)=19.67 and $p < .001$; Internet (non-social media) with X2 (1, N=667)=44.99 and $p < .001$; cellphone weather apps with X2 (1, N=667)=39.141 and $p < .001$; and social media other than Twitter with X2 (1, N=667)=50.85 and $p < .001$. Compared to non-Twitter users, a greater percentage of Twitter users reported receiving weather-related information via the four modalities listed above. Within the full sample (including both Twitter users and non-Twitter users), there exists a difference in ways of receiving information when taking into account the loss of electrical power. Television was the major source of information for 82% of respondents with power, which is consistence with past findings (Feldman et al., 2016; Burke, Spence, & Lachlan, 2010). Internet and radio were the other main sources of receiving storm-related information for 32% and 29% of the sample respectively (Figure 2.4). Only 2% of respondents received the information through Twitter. Similar results were observed in a study by Feldman et al. (2016), confirming the low usage of Twitter (2.3) in communications related to flood risk for 164 residents of Newport Beach, California. However, Twitter was more desired as a medium of future communications among these residents. While Feldman et al. (2016) did not consider the impact of power outage on the use of Twitter, our analysis shows how the respondents who lost power during Hurricane Sandy reported having received storm-related information (Figure 2.5). In our study, individuals who lost access to electrical power, moved away from their reliance on television for receiving weather-related information and toward a reliance on radio, telephone, cellphones (through both text messages and weather apps), and the Internet, including social media. Interestingly, 18% of respondents without power used Twitter as their information source.

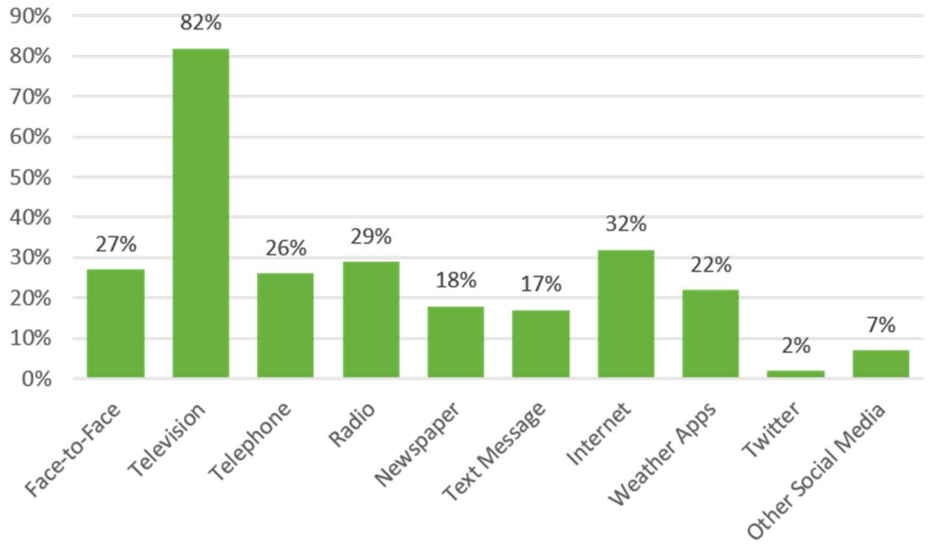


Figure 2. 4. Information Sources for Respondents with Power (n=205).

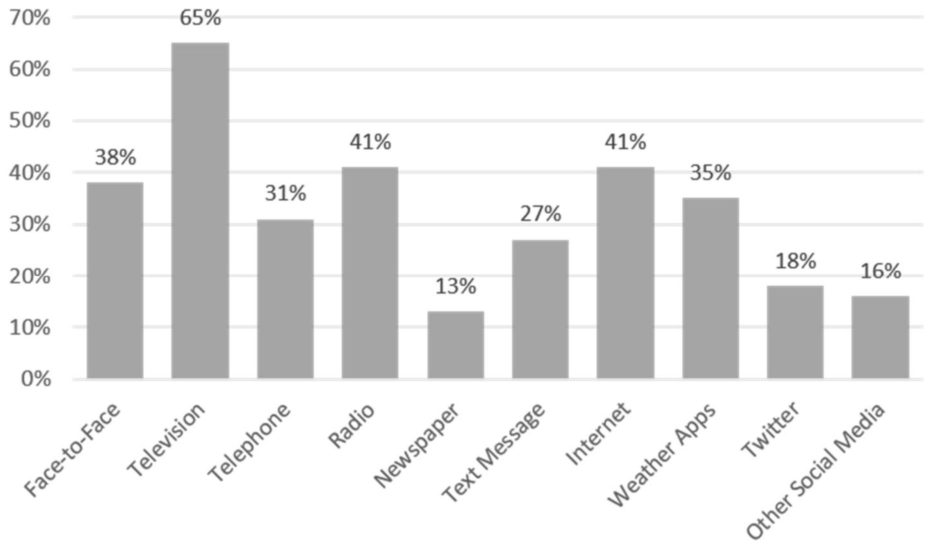


Figure 2. 5. Information Sources for Respondents without Power (n=414).

Respondents were also asked whether they received information from any of seven different sources including friends, family, household member, local news, national news, federal agencies, and state agencies. Figures. 2.6 and 2.7 represent the sources of weather-related information for Twitter users and non-Twitter users, respectively. The positive impact of news media engagement on Twitter activities during disruptive events has been identified in literature (Saleem, Xu, & Ruths, 2014b). While both Twitter users and non-Twitter users in our study relied on local and national news at about the same frequency, Twitter users reported having received information from family, friends, and government agencies at a higher rate than non-Twitter users. This suggests Twitter users depend more on information accessed via their family, friends (Abbas, Bayat, & Ucan, 2018), and government agencies when compared to non-Twitter users. One potential application of Twitter data, therefore, is that government agencies can utilize Twitter during emergency situations and, by doing so, reach at least a subset of the population most at risk during weather-related emergencies. However, our social network analysis shows (see section 2.5.3.1) the bi-directional communication does not usually happen during emergencies.

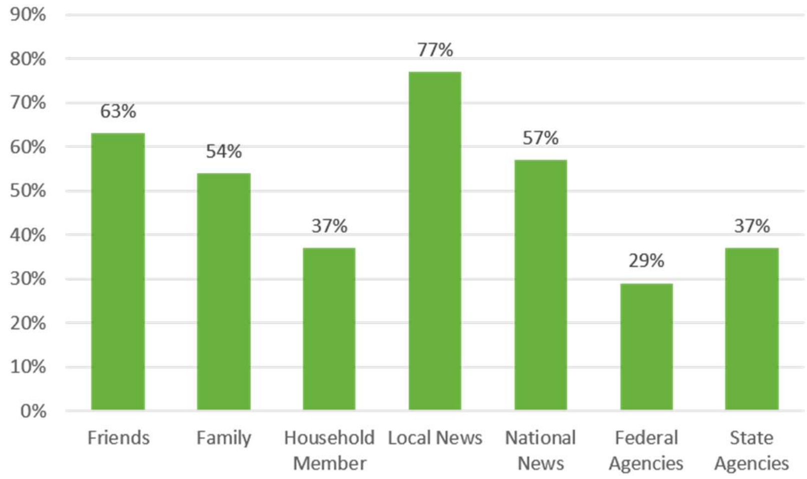


Figure 2. 6. Weather-Related Information Sources for Twitter Users (n=266).

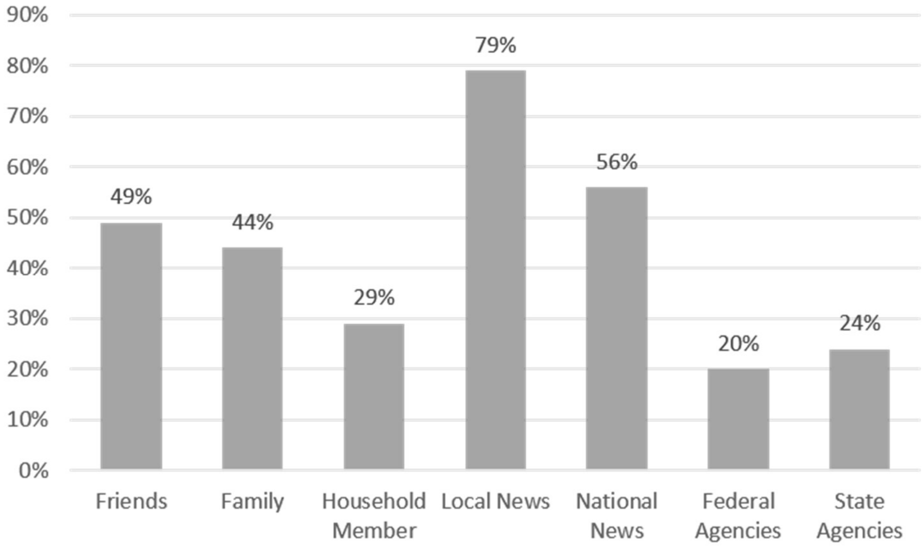


Figure 2. 7. Weather-Related Information Sources for Non-Twitter Users (n=420).

A secondary analysis relating to this research question examined what sources the Twitter users followed on Twitter (Figure. 2.8). While Twitter users followed a myriad of weather-related sources, the three top sources were: 1) Local Television News, 2) The National Weather Service, and 3) The Weather Channel. Approximately 59% of Twitter users in the survey reported that they shared weather-related information via Twitter during Hurricane Sandy. Figure. 2.9 illustrates the types of information these users shared. The data indicate that photographs were the most frequently shared form of information (62%) by Twitter users, followed (in frequency) by personal experiences (56%), and information about storm damage (53%). It is not necessarily the case that the categories of information Twitter users shared are mutually exclusive, meaning respondents may have posted a tweet sharing their personal experiences with the storm and attached an image to that tweet. Regardless, these findings illustrate Twitter's utility as an image-sharing platform.

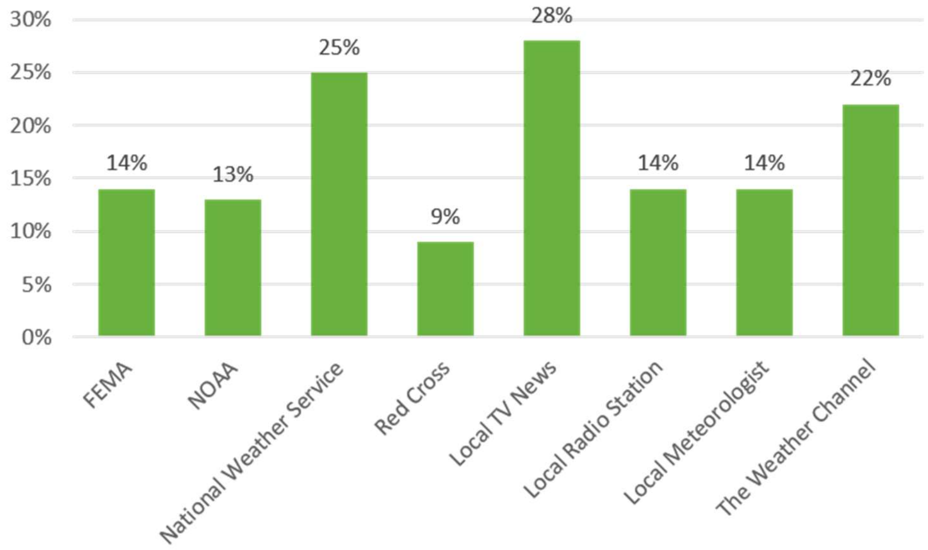


Figure 2. 8. Weather-Related Twitter Accounts Followed by Survey Respondents (n=266)

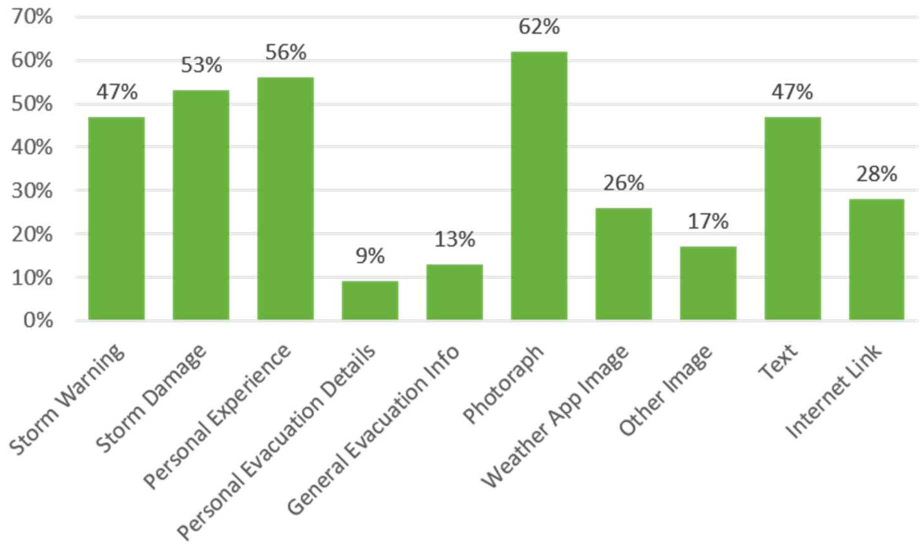


Figure 2. 9. Types of Information Twitter Users Shared During Hurricane Sandy (n=156).

2.5.2 Research Objective 2: To Examine Twitter Users' Involvement in Discussions about Hurricane Sandy and the Content of Messages Shared Before, During, and After the Hurricane

Twitter data analysis addressed the following three research questions: (1) How did the participation rate of Twitter users in discussions about Hurricane Sandy change during the three phases of analysis? (2) What were the top words posted by Twitter users and how did they change during the three phases? (3) How did the positive/negative sentiments of shared messages change over the three phases?

2.5.2.1 Temporal analysis

The temporal analysis of the keyword and the geo-tagged datasets shows that the results are in line with past research that use of social media in a disaster starts very early (Crooks, Croitoru, Stefanidis, & Radzikowski, 2013) and reach its peak mostly while it is happening (Caragea, Squicciarini, Stehle, Neppalli, & Tapia, 2014). Analysis of the keyword dataset (Figure 2.10) shows that there was a substantial increase in the number of messages and unique users contributing to the hurricane discussions from Oct 26, 2012 with the highest peak occurring on

Oct 29, 2012 (approximately 237 K unique messages being shared per hour) at 6:00pm EST. At the time Hurricane Sandy made landfall (8:00pm EST in Atlantic City), approximately 223 K unique messages were being shared across the network by 187 K unique users per hour. In the following days, both during- and post-Hurricane Sandy, the number of Twitter messages along with the number of users decreased over time. The post-hurricane phase revealed a larger number of messages being shared across the network in comparison to the pre-hurricane period. This is due to large frequency of relief-related tweets in the aftermath of landfall. For example, the increase in the Twitter users and tweets on November 2 was related to NBC's live telethon, ‘Hurricane Sandy: Coming Together’ encouraging Twitter users to live tweet using the hashtag #SandyHelp (Murthy & Gross, 2017). The general patterns observed in this study are in consistent with other Sandy-related studies (Murthy & Gross, 2017; Chatfield et al., 2014) showing the accuracy of our collected data and analysis.

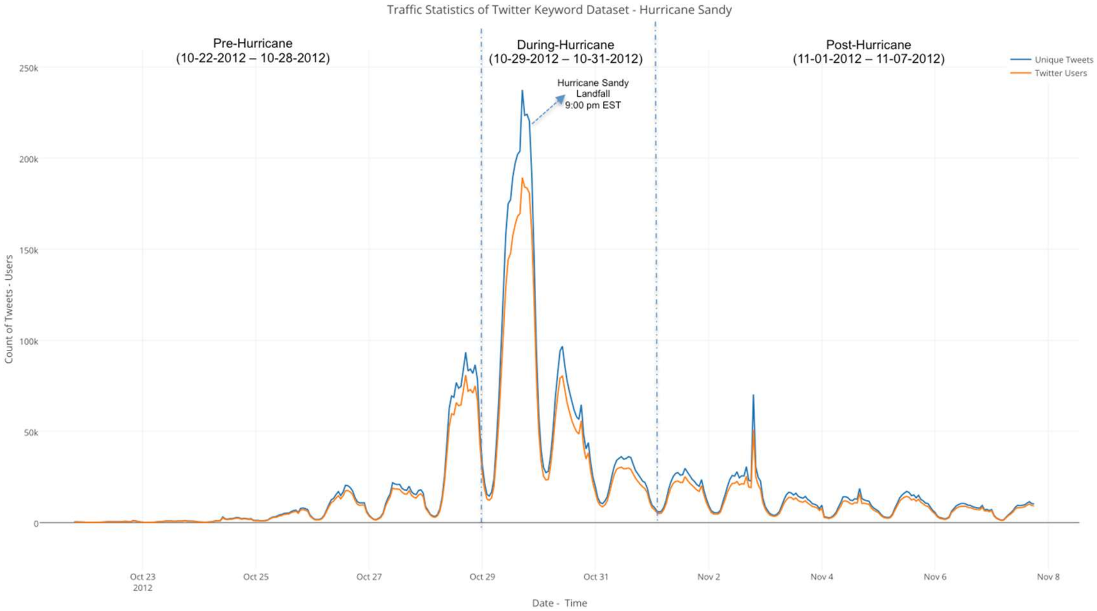


Figure 2. 10. Temporal Analysis of Keyword Dataset Aggregated by Hour.

Figure 2.11 shows the temporal dynamics of the geo-tagged messages and their corresponding users collected from the New York, New Jersey and Connecticut area. In comparison to the keyword dataset, the geo-tagged traffic shows a large number of peaks in all three phases of the hurricane. During the pre-hurricane phase, the number of tweets shared from the locations decreased on Oct. 28 and then gradually rose to the peak during landfall at 8:00pm EST on Oct. 29. The traffic then gradually decreased over the following days and peaked again on Nov. 7, which was a result of President Obama's re-election. These messages provide valuable information from individuals residing in the hurricane-affected areas. More specifically, the messages contain ed firsthand information about the hurricane, along with disaster-related images taken in real-time (Figure 2.12). A large number of posts contained weather-specific photographs showing the intensity of the hurricane in real-time, along with images of damage and flooding through which researchers and emergency managers can retrieve information to help identify storm damage and plan relief efforts. With the growing number of Twitter users particularly during power outages as our study shows, this information can be useful for a better disaster management.

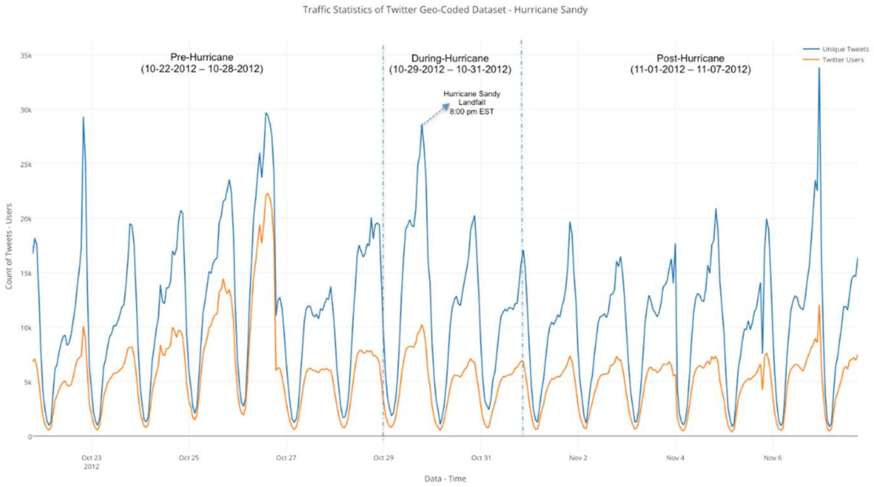


Figure 2. 11. Temporal Analysis of Geo-Tagged Tweets Aggregated by Hour.

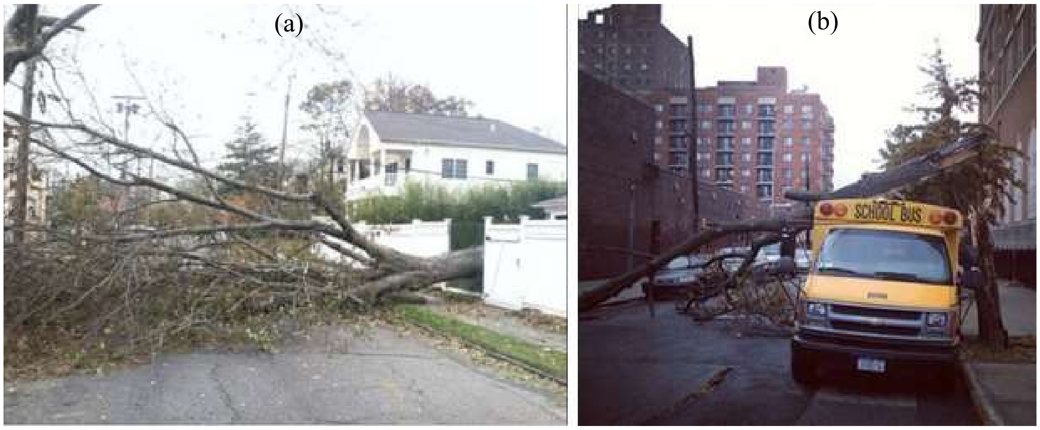


Figure 2. 12. Photos in New Jersey (a) and New York (b) Shared by Two Twitter Users on October 30.

2.5.2.2. Klout analysis

The analysis of the aggregated Klout scores shows a decrease in the average score per hour nearing landfall (Figures 2.13 and 2.14). This indicates an increase in participation of the general population (lower Klout Scores in comparison to influential users) in the discussions leading up to, and following, Hurricane Sandy. While influential users with high Klout scores also participated in the discussions, the general population were more active in sending messages that were being shared across the network. The analysis also shows that for the geo-tagged dataset, users have a much lower average Klout score than the keyword dataset users. This may indicate that the majority of the general population have their geo-location services enabled on their mobile devices. While users with higher scores have been identified as more effective in spreading the information (Rao et al., 2015), geo-tagged tweets shared by general public might be more valuable for identifying the most affected people and areas for a better disaster management.

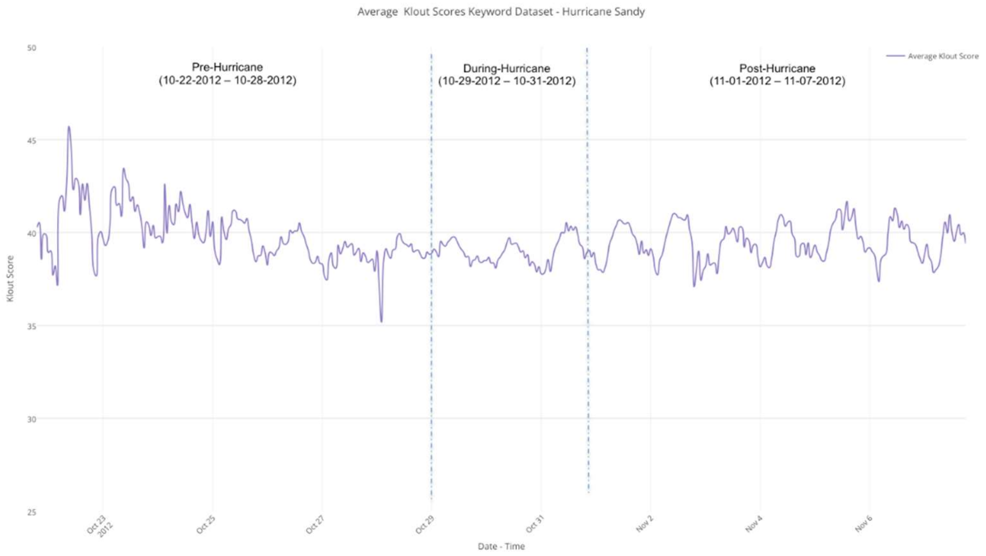


Figure 2. 13. Average Klout Scores Aggregated by Hour in Keyword Dataset.

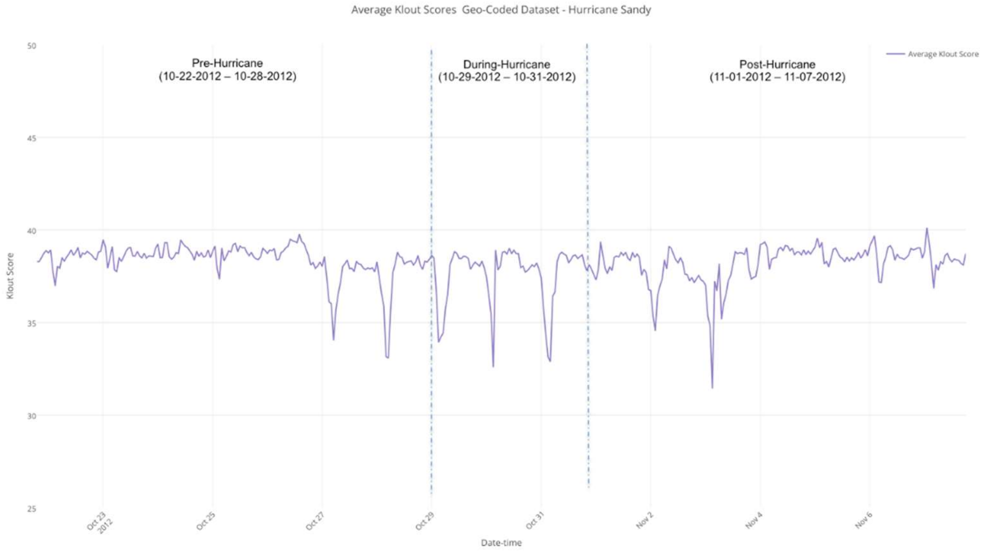


Figure 2. 14. Average Klout Scores Aggregated by Hour in Geo-Tagged Dataset.

2.5.2.3 Text analysis

The word cloud analysis shows clear differences among the words used before, during, and after the hurricane, a result confirmed in other disaster-related studies (Jooho Kim et al., 2018). The level of hazards and risk, and trend of the incident can be identified through word analysis (Jooho Kim et al., 2018). In the keyword dataset (Figure 2.15), analysis of the word distributions reveals that the most frequently occurring words in the pre-hurricane phase were: *Sandy, Hurricane, Frankenstorm, Storm, New York, Coming,* and *Tomorrow*. In comparison, the most frequently occurring words during hurricane were: *Sandy, Hurricane, Power, HurricaneSandy, Safe, Stay Safe, East, Prayer,* and *Good*. In the post-hurricane phase, the most frequently occurring words were: *Help, Relief, Sandy, Hurricane, New York, SandyHelp, Aftermath,* and *Power*. The analysis shows the transition of discussion across the different phases of the hurricane from people advising and spreading the news of the hurricane → Twitter users being concerned about the well-being of their friends and followers → relief and rescue efforts in the aftermath of the hurricane. The results are paralleled with another study by Spence et al., (Spence, Lachlan, Lin, & del Greco, 2015) examining the content of tweets during Hurricane Sandy. The discussion of power outages also occurred in the during- and post-phases of the hurricane. The geo-tagged dataset presents similar results with one key difference in the post-hurricane phase, where the discussion was more of an everyday social interaction along with some discussions on President Obama's re-election (Figure 2.16).

The word co-exsistence analysis shows that the words *hurricane* and *sandy* were the top-ranking words in the pre-phase (Figure 2.17). They were connected by less occurring words such as *frankenstorm, storm, hurricanesandy, school,* and *monday.* The during-phase (Figure 2.18) shows a substantial change in the words that were tweeted. Words that formed clusters included: [*stay, safe, strong*], [*friends, family, share*], [*New, York, Jersey, Atlantic*], [*prayers, thoughts*], and [*power, out, home*]. The increasing use of more emotional tweets in this phase might be a good mechanism to release anxiety or stress, return people to a normal state, and provide a sense of connection among those who experience the similar situation (Lachlan, Spence, Lin, & Del Greco, 2014). In the post-phase (Figure 2.19), clusters related to the aftermath [*power, still, out*], relief [*food, water*], [*help, relief, sandyhelp*] and donation [*redcross, donate*] formed the main topics of discussion.

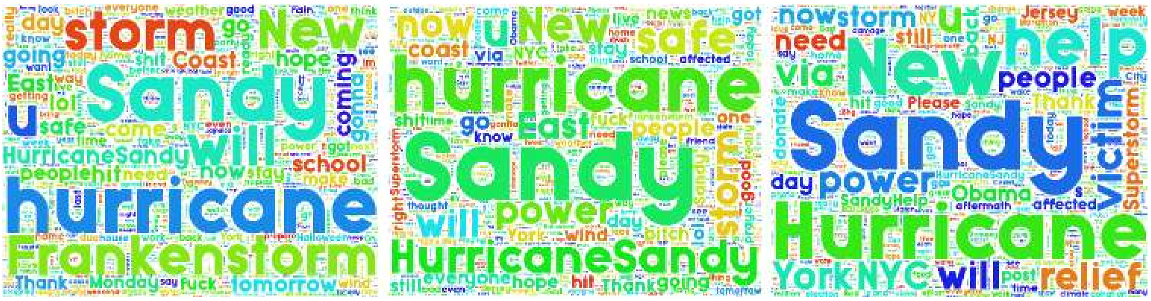


Figure 2. 15. Word Clouds in Pre, During and Post Hurricane Phases (Keyword Dataset).

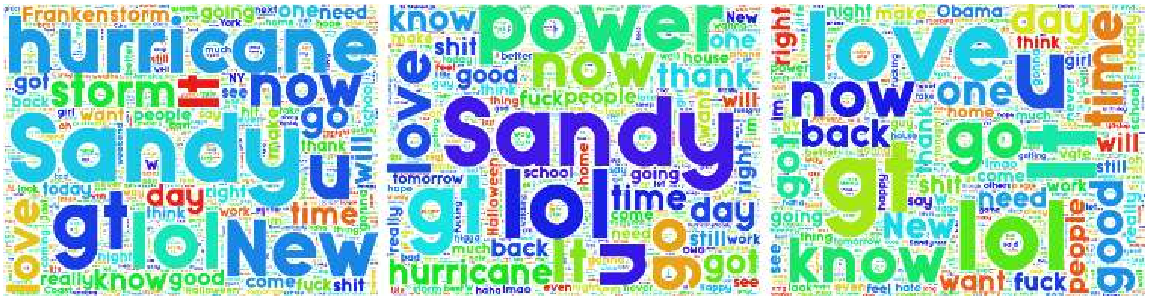


Figure 2. 16. Word Clouds in Pre, During and Post Hurricane Phases (Geo-Tagged Dataset).

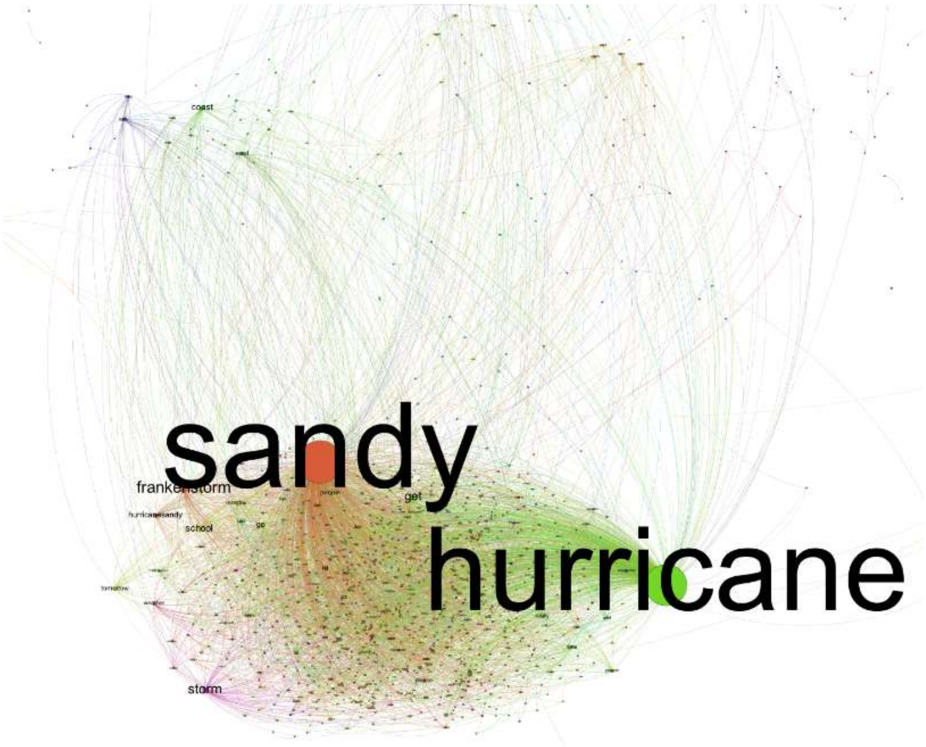


Figure 2. 17. Word Co-Existence: Pre-Hurricane Phase (Keyword Dataset).

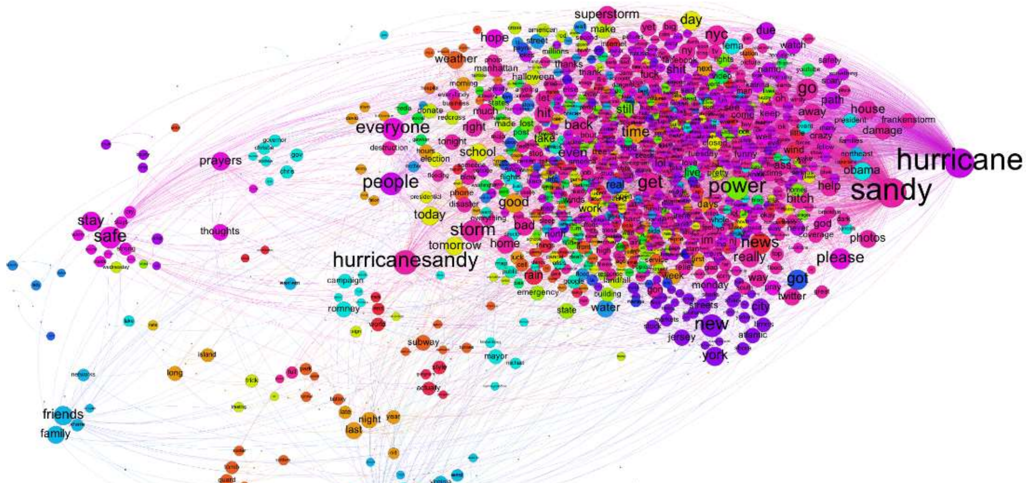


Figure 2. 18. Word Co-Existence: During-Hurricane Phase (Keyword Dataset).

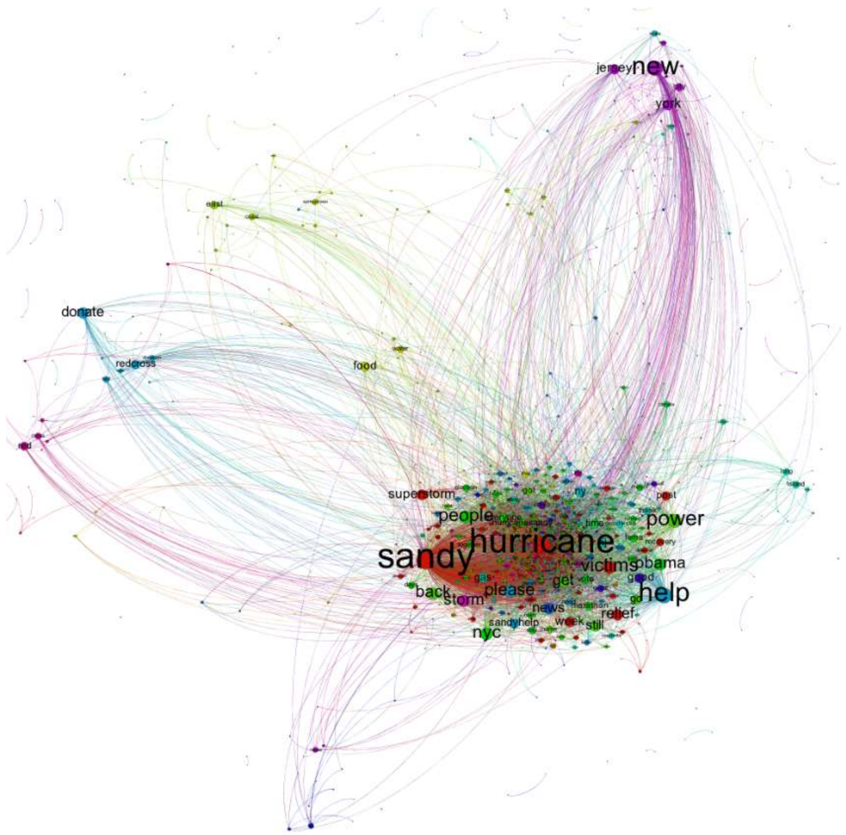


Figure 2. 19. Word Co-Existence: Post-Hurricane Phase (Keyword Dataset).

With the sheer volume of data shared during emergencies, one key problem is how to select proper information at different stages. Looking for keywords such as *sandy* or *hurricane* will result in a large dataset that might not be helpful. A closer look at the frequency and coexistence of words can help us to identify specific keywords to search for at different stages of emergencies. For example, the keywords *please*, *need*, *flooding*, and *power* had a higher frequency in the during-phase of Hurricane Sandy (Figure 2.20). The word *victim*, on the other hand, had a higher frequency in the post-phase. These keywords might be useful to locate the event and people in need during disasters. Another useful information are Tweets containing the word *evacuate*. These tweets were mostly started on 28th and increased sharply during 29th when the mandatory evacuation order was announced. These data in combination with geo-tagged information can help in understanding the evacuation behavior of residents.

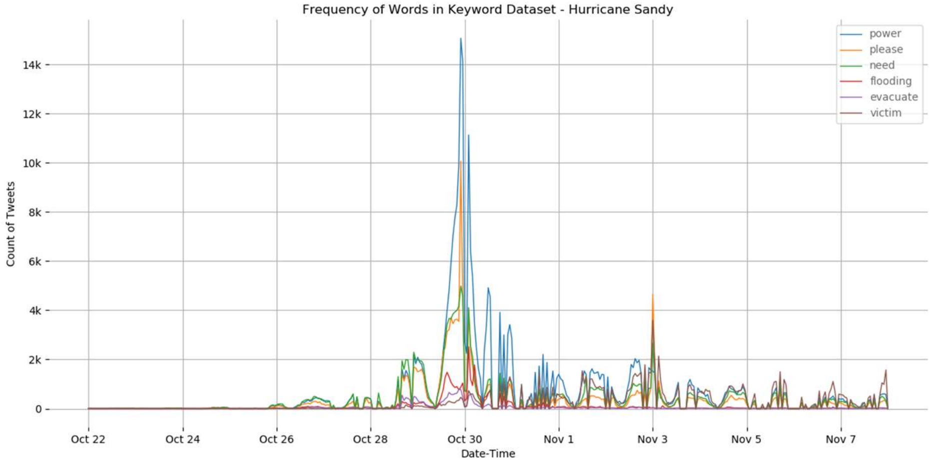


Figure 2. 20. Frequency of Words Aggregated by Hour (Keyword Dataset).

The coexistence analysis also revealed valuable information. The word *power*, for example, coexisted with words such as *outages*, *knock*, *lost*, *lose*, *cuts*, *downed*, *dead*, *nuclear*, *plants*, *customers*, *homes*, and *still*, in the during-phase of Hurricane Sandy. These words might be appropriate search keywords for power-related companies to identify affected areas. In the post-phase, the words *food*, *water*, *shelter*, *clothes*, *distribution*, *drive*, *truck*, and their co-existence with words such as *victims* and *need* not only show useful keywords to search for, but also are valuable for disaster-related agencies to identify victims and provide their urgent needs in a timely manner. The co-existence of the words *gas*, *lines*, *stations*, *situation*, *problems*, *shortage*, and *long* shows other useful keywords that can help to identify gas-related problems in specific areas.

2.5.2.4 Sentiment analysis

Sentiment analysis of different phases of the study (Figures 2.21 and 2.22) reveals that the messages shared by Twitter users were more negative in the during-hurricane phase of the study. The pre-hurricane phase of the study in both datasets displayed a more positive sentiment, while gradually decreasing toward the mid-point of the during-hurricane phase (lowest sentiment average scores) and then increasing in the post-hurricane period. This shows the dynamics of Twitter users who were posting messages with an increasingly negative attitude to Hurricane Sandy at the peak of the storm. The comparison of the keyword and geo-tagged datasets also reveals that the sentiment of Twitter messages originating from the Hurricane Sandy areas (geo-tagged dataset score range between −0.2 and +0.5) are more negative than the keyword dataset (score range between +0.1 and +0.5) suggesting the influence of distance (from disaster area) on sentiments of tweets. While sentiment analysis is a valuable approach to detect and locate disasters (Kryvasheyeu, Chen, Moro, Van Hentenryck, & Cebrian, 2015), it is not fully used by authorities. This is due to the inability to efficiently sort and categorize the sheer volume of data generated during disasters (Caragea et al., 2014). Our analysis is a confirmation of the concept that for a better utilization of social media data during natural disasters and this type of analysis, authorities should mostly use the negative tweets in the during-phase of disaster, when more people use social media to communicate their needs. Selecting these negative tweets would be more helpful to detect and locate people at risk in a timely manner rather than focusing on all tweets. Evaluating negative tweets in time and space and integrating them into a system that use various modalities such as text and network analysis remains for future research

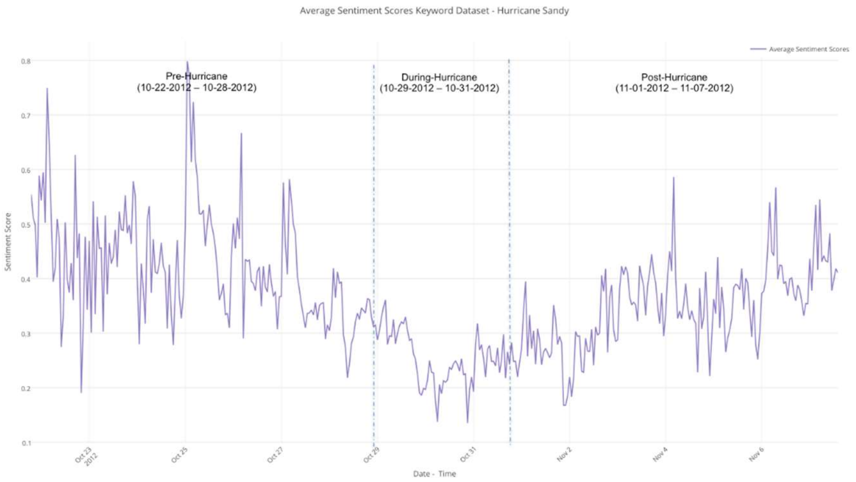


Figure 2. 21. Average Sentiment Score Aggregated by Hour Before, During and After Hurricane (Keyword Dataset).

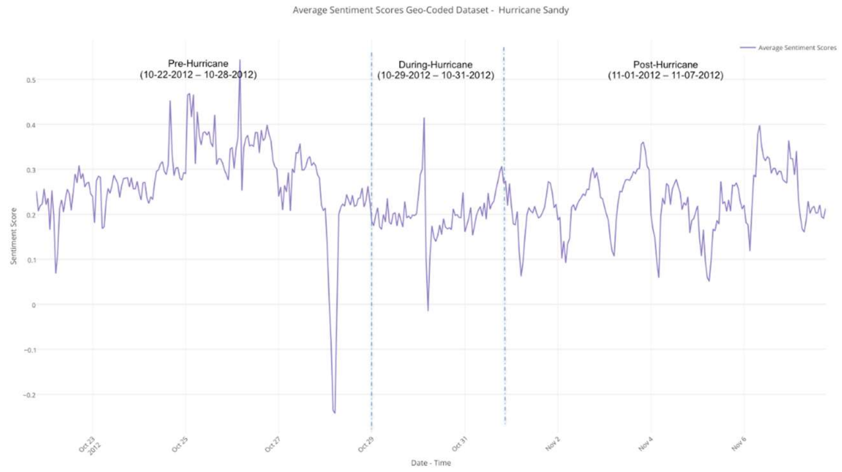


Figure 2. 22. Average Sentiment Score Aggregated by Hour Before, During and After Hurricane (Geo-Tagged Dataset).

2.5.3 Research Objective 3: To Examine the Social Network Structure of Twitter Users Before, During, and After the Hurricane

Network analysis of the users addressed the following three research questions: (1) How did the network structure of Twitter users evolve during the three phases of Hurricane Sandy? (2)

How did the users' communities form during the three phases? (3) Who were the key influencers across the networks?

2.5.3.1 Social network analysis

We used users' networks to analyze the interconnectivity of Twitter users during each phase of the hurricane. Network data statistics are shown in Table 2.3. The highest numbers of nodes and edges were observed in the during-hurricane phase representing a larger number of active users communicating during the hurricane. The larger network diameter in the during-phase also supports the analysis with larger participants increasing the size of the connected network. Average degree and weighted degree are larger in the post-hurricane phase indicating stronger bi-directional communication in the relief efforts. In-degree (the number of incoming edges) and out-degree (the number of outgoing edges) centrality measures represent the prominent and influential users in a network respectively (Hanneman & Riddle, 2005).

Table 2. 3 Network Statistics of Twitter Users (Keyword Dataset).

	Pre-hurricane	**During-hurricane**	**Post-hurricane**
Network Type	Directed	Directed	Directed
Nodes	10,788	50,607	32,879
Edges	36,269	234,000	176,396
Average Degree	3.362	4.624	5.365
Average Weighted Degree	4.961	6.269	8.284
Network diameter	17	30	18
Graph Density	0.000	0.000	0.000
Average Path Length	5.472	6.683	5.605
Modularity	0.571	0.502	0.521
Average Clustering Coefficient	0.057	0.045	0.069

In our study, news agencies, political figures, weather, and other disaster-related agencies and organizations (e.g., FEMA, Red Cross) were the users with top in-degree centrality (Table 2.4). However, the low out-degree measures of these users show the one-way communication between these users and general public. While many people mentioned disaster-related agencies such as FEMA, these users were not much responsive. Similar results were observed for the users with the highest eigenvector centrality (Table 2.5). The assumption behind the eigenvector centrality is that the centrality of a node is proportional to the sum of the centralities of its neighbors, therefore it describes how well the node in the network is connected to other well-connected nodes (Oliveira & Gama, 2012). This suggests the central/influential roles of identified users in the overall structure of the network (Hanneman & Riddle, 2005) as potential sources of disaster-related information and social engagement. However, the users spreading the information were mostly among public figures as shown by out-degree measures (Table 2.6), a result observed in other disaster-related studies (Jooho Kim et al., 2018). These public figures were engaged with approximately 100–200 other users, a much lower number compared to the number of users who mentioned top in-degree users.

Table 2. 4 Users with Top In-Degree Measure and Their Associated Out-Degree.

Pre-hurricane			**During-hurricane**			**Post-hurricane**		
	In	Out		In	Out		In	Out
MikeBloomberg	599	13	GovChristie	4312	0	GovChristie	5140	27
fema	467	26	MikeBloomberg	2636	20	MikeBloomberg	4365	14
twc_hurricane	458	47	NYGovCuomo	2468	55	RedCross	3731	3
NHC_Atlantic	420	0	CoryBooker	2116	93	CoryBooker	3299	201
GovChristie	396	7	NYCMayorsOffice	1953	20	nytimes	3125	0
NYCMayorsOffice	396	18	twc_hurricane	1933	117	ABC	2961	0
weatherchannel	384	17	RedCross	1929	11	NYGovCuomo	2798	67
AP	376	0	BarackObama	1856	0	FoxNews	2791	0
CoryBooker	367	67	AP	1653	3	CBS	2428	0
JimCantore	359	5	cnnbrk	1499	0	NPR	2328	0

Table 2. 5 Users with Top Betweenness and Eigenvector Measures

Pre-hurricane		During-hurricane		Post-hurricane	
Betweenness	**Eigenvector**	**Betweenness**	**Eigenvector**	**Betweenness**	**Eigenvector**
twc_hurricane	MikeBloomberg	FDNY	GovChristie	CoryBooker	GovChristie
CoryBooker	NHC_Atlantic	twc_hurricane	MikeBloomberg	ConEdison	MikeBloomberg
WSJweather	fema	CoryBooker	NYGovCuomo	fema	RedCross
garytx	twc_hurricane	NYGovCuomo	NYCMayorsOffice	MikeBloomberg	CoryBooker
NYGovCuomo	weatherchannel	RedCross	CoryBooker	FDNY	NYGovCuomo
fema	NYCMayorsOffice	AntDeRosa	RedCross	NYGovCuomo	nytimes
RyanMaue	RedCross	rqskye	twc_hurricane	GovChristie	ABC
weatherchannel	JimCantore	fema	BarackObama	blogdiva	FoxNews
JimCantore	NOAA	ConEdison	AP	RedCross	NYCMayorsOffice
USCG	GovChristie	granthansen	fema	JCP_L	BarackObama

Interestingly, disaster-related agencies such as FEMA were not among the top out-degree users, while the public expectation is to receive more information from these users. Betweenness centrality measures the extent to which a node in the network lies between other nodes (Oliveira & Gama, 2012). Our results for the top betweenness centrality (Table 2.5) include political and public figures, weather agencies, and other disaster-related agencies and organizations suggesting their role as bridges in the communication network. Users with high betweenness centrality in a network are also called gatekeepers since they control how information flow between communities (Oliveira & Gama, 2012). Public and political figures can play a significant role during disasters. Authorities can benefit from connecting to these users to spread relevant information and receive more information on the people's need, but our analysis did not find these connections.

Table 2. 6 Users with Top Out-Degree Measure and Their Associated In-Degree.

Pre-hurricane			During-hurricane			Post-hurricane		
	Out	In		Out	In		Out	In
editchick	102	0	rightnowio_feed	203	10	CoryBooker	201	3299
weatherplaza	95	6	SeanPCollins	183	16	farside314	181	7
MarnieTWC	73	25	weeddude	180	101	BlondeVelvet	163	6
JustCouch	71	5	ScottBeale	150	64	blogdiva	158	78
CoryBooker	67	367	JustCouch	146	0	wishuponahero	154	17
blogdiva	64	23	ninatypewriter	143	53	GrandmaJer_ETSY	134	0
sahnetaeter	62	6	DAKGirl	142	11	HealthcareWen	115	40
HumanityRoad	58	30	farside314	137	0	EarlyShares	114	3
trdelancy	58	6	azipaybarah	132	80	ConEdison	104	1770
weeddude	57	14	SustainablDylan	122	0	SINYCliving	100	114

We identified and visualized the user community clusters using the Louvain modularity measure. The nodes represent Twitter users and the lines between the nodes represent the connectivity by the attribute of mention. The size of the nodes (and the size of username) represents the number of times a particular user was mentioned. The varying colors of the nodes and the edges represent the various clusters. In the pre-hurricane phase (Figure 2.23), political Twitter users such as, MikeBloomberg, NYGoverner, NYCMayorsOffice, CoryBooker clustered together. Whereas news agencies (NHC_Atlantic, breakingstorm, wunderground, twc_hurricane, weather_channel, ABCnews) and federal agencies (FEMA, NOAA) each form separate clusters. RedCross and GovChristie have their own influence groups, which are separate from the other clusters. During hurricane (Figure 2.24), most of the users clustered together, suggesting a dense network of communication with information being shared in a bi-directional model across the platform. GovChristie achieved the highest number of mentions, followed by other political Twitter users (BarakObama, NYCMayorsOffice, NYGovCuomo, MikeBloomberg, CoryBooker). FEMA and RedCross gravitated closer to news agencies such as HuffPost, nytimes, AP, cnnbrk. The formation of a blue circular cluster (on the right side of the data structure) proved to be noise unrelated to the hurricane. This cluster represents a high volume of communication between two

United Kingdom musical bands and their producers' discussions about a song titled "Sandy." In the post-phase of the hurricane (Figure 2.25), users clustered near the center of the graph, suggesting that most of the centroid users where highly connected to other users in the network. RedCross was highly mentioned as a result of its involvement in the relief efforts.

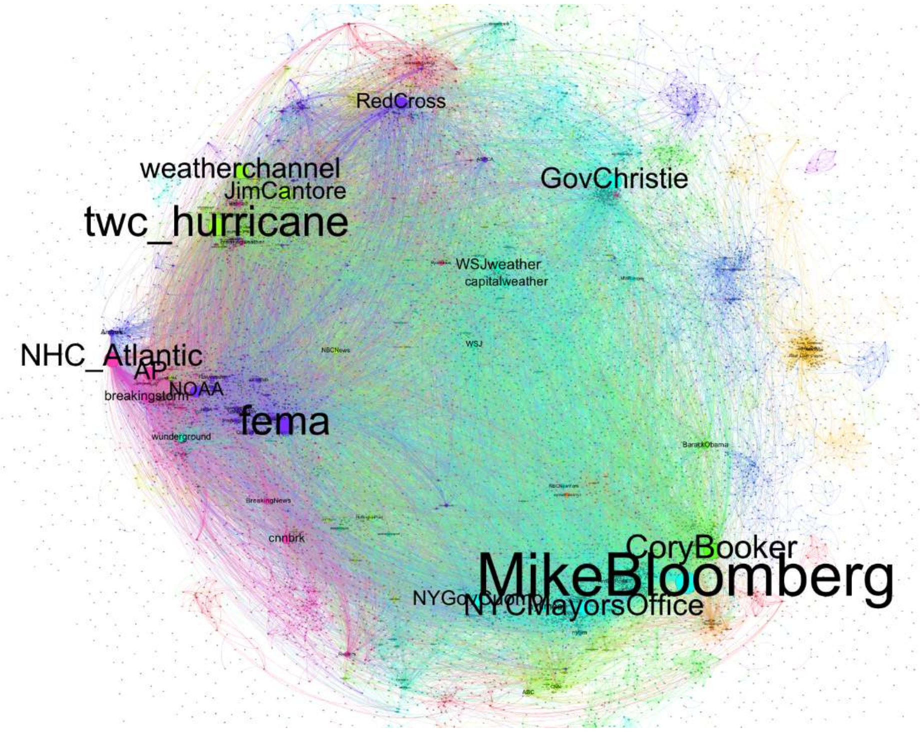


Figure 2. 23. User Network Analysis: Pre-Hurricane Phase (Keyword Dataset).

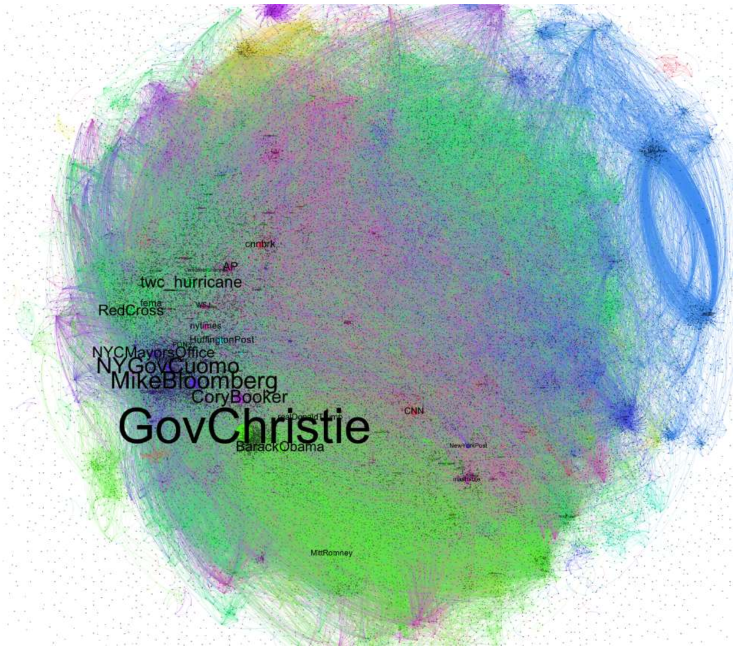


Figure 2. 24. User Network Analysis: During-Hurricane Phase (Keyword Dataset).

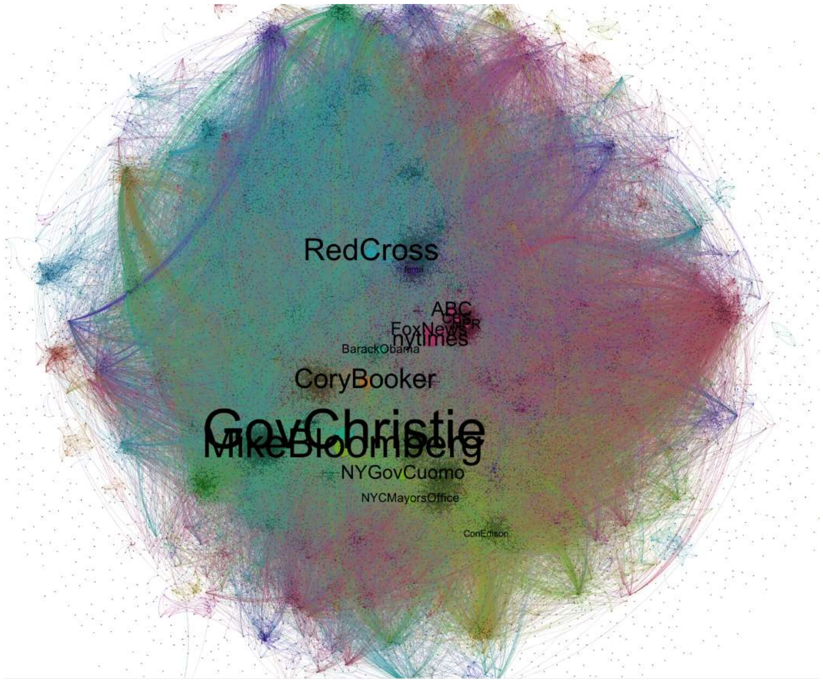


Figure 2. 25. User Network Analysis: Post-Hurricane Phase (Keyword Dataset).

2.6 Discussion and Conclusions

We examined the usage of social media during Hurricane Sandy from both survey and communication perspective. Research has validated the prominent role of social media, particularly Twitter as a source of information (Landwehr, Wei, Kowalchuck, & Carley, 2016; Westerman, Spence, & Van Der Heide, 2012). Limitations of other media sources in providing useful information during the disaster is one of the reasons for using social media (Sutton et al., 2008). The results of our study indicate that Twitter can serve as a valuable medium for communication through sharing texts and photos during weather-related emergencies, especially during power outages. Our findings also indicate that Twitter users generally receive emergency information from various sources at higher rates than non-Twitter users. This is due to the easier access of Twitter users to prominent individuals, organizations, and agencies who share weather-related information as our social network analysis reveals. Temporal analysis of tweets showed that use of Twitter during Hurricane Sandy started very early and continued in the aftermath of disaster. With the change of discussions in different phases of hurricane from people advising and spreading the news, to concerns about the well-being of their friends and followers, and to relief and rescue efforts in the aftermath of the hurricane. The emergency officials can use such messages as a valuable source of information from individuals residing in the hurricane-affected areas. Lower sentiment scores (negative polarity) of tweets originating from the affected areas also revealed the influence of distance on Twitter messages. Authorities should mostly use the negative tweets in the during-phase of disaster to detect people at risk and their immediate needs.

Our major aim was to understand the communication dynamics through different modalities including temporal, text, user, sentiment, and social network analyses and identify the most important information that can be derived from twitter during disasters. Government/ relief/emergency agencies can utilize such information to reach and support at least a subset of the population most at risk. While government agencies are among the prominent Twitter users during disasters, they primarily rely on one-way communication rather than engaging with their audiences (Waters & Williams, 2011), a result confirmed by our network analysis. A major advantage of social media over traditional media sources is its potential for a bi-directional communication where public could provide the agencies with useful information for disaster management (Lachlan et al., 2014). However, this was not the case during hurricane Sandy.

With the increasing frequency of extreme weather events, Twitter users' networks along with tweets' texts and geolocation information provide real-time data that can equip emergency management tools. Our findings illustrate Twitter's utility as an image-sharing platform that might be useful for developing applications to identify relevant images for disaster response using the location information embedded in the messages. However, the major challenge is the quality assessment of such data (Eshghi & Alesheikh, 2016). It is important to evaluate the relevancy and credibility of data and select only tweets related to the event. While current research is mostly focused on keyword filtering, developing a new approach integrating various data types is an important step for future research to utilize social media successfully in natural disasters. In our future research, we will develop an automated data-learned model to filter geocoded images shared across the Twitter platform by leveraging a multi-model that analyzes geospatial, image, user and text data.

Twitter is a useful medium through which individuals can communicate during weather-related emergencies; a valuable source for researchers to better understand these communications; and a useful way for agencies to communicate with individuals at risk. However, various barriers (e.g., lack of resources, lack of organizational support) to the use of social media analytic tools in organizations need to be first addressed to facilitate disaster management efforts (Anson, Watson, Wadhwa, & Metz, 2017). With the difficulty of sorting and locating relevant information for public (Lachlan et al., 2014), the government agencies should facilitate the process by developing automated emergency management tools that extract and analyze information shared on social media. These tools can also benefit from social network analysis to help authorities in accelerating information diffusion during disasters. Future studies should examine other social media platforms in a multi-case approach to increase the effectiveness of social media in disaster management.

CHAPTER III

SPACE, PLACE, AND FLOWS: SPATIOTEMPORAL ANALYSIS OF TWITTER DURING NATURAL DISASTERS [2]

3.1 Introduction

Space and place have long been core topics in academic literature (Yang et al., 2016). Geographic information systems (GIS) has been dominated by the concept of space—represented by Cartesian coordinates (Sui & Goodchild, 2011). Place, however, has not been thoroughly formalized in GIS and it has been often used interchangeably with the term space (Sui & Goodchild, 2011). The emergence of information technology in recent years has provided new opportunities to study human behavior and activities from a place perspective—a "bottom-up" view that captures the local environment and human activity in a qualitative manner (Yang et al., 2016). Geospatial big data such as social media have enabled a number of users to share information at any time in any location (Zou, Lam, Cai, & Qiang, 2018). The real-time characteristic of social media and their rich content not only have made them invaluable sources for exploring the place concept in the GIS domain (Yang et al., 2016), but also suitable crowdsourcing platforms to become aware of any event in those places.

This enhanced awareness through social media has become of particular importance to both scholars and emergency management authorities especially due to the shortcoming of traditional data in observing human behavior during emergencies (Zou et al., 2018). Understanding the emergency situations is an important step for providing relevant information to population at risk and reducing human and economic impacts (Martínez-Rojas, Pardo-Ferreira, &

[2] Pourebrahim, N., Sultana, S., & Mohanty, S. (2019). *Cities.* In review.

Rubio-Romero, 2018). Therefore, appropriate use and analysis of social media data is an important task for achieving the full potential of social media during emergencies (Lindsay, 2011). However, big data analyses have been criticized for their failure to address methodological issues; their non-theoretical approaches; and their claims on transcending the need for any domain expertise in the analyzed subjects (Shelton et al., 2014). To address these critics and enhance public awareness of a situation in emergencies, one need to first address questions such as: How to extract useful information from social media to understand human activities and events? And how this information can be linked to the theoretical concepts such as space and place in the GIS domain?

Existing studies have used different spatial techniques to improve situational awareness (Wang, Ye, & Tsou, 2016). The combination of geospatial information and content of social media has gained more attention recently (i.e Wang & Ye, 2018; Kryvasheyeu et al., 2016). However, the integration of the qualitative text data with spatio-temporal data to identify specific events during disasters from a place perspective has less been investigated. Moreover, study of people's movement during such events is another aspect less explored. Inclusion of collective movement patterns in the situational awareness process and understanding their correlation with the embodied social processes can help in a better disaster management. To address these challenges, this paper examines social media analysis of disasters in the GIS domain. The objectives are: 1) to integrate spatio-temporal data with topic modeling of qualitative text data to identify places of events during Hurricane Sandy; 2) to identify places created by out-flow movements during Hurricane Sandy; and 3) to explore the association between out-flow movements and the underlying social/physical vulnerability of places.

The rest of paper is organized as follows: section 2 provides a brief review of concept of place and situational awareness. The study area, data, and methodology are discussed in section 3. Results are discussed in section 4, followed by conclusion and future work in the last section.

3.2 Related Work

3.2.1 Space, Place, and Scale

Space and place are the fundamental concepts of geography. While space is related to an abstract "top-down" view, place is more concreate with a "bottom-up" perspective (Yang et al., 2016). Places are the spaces that people are attached to it in one way or another, a meaningful location as stated by Cresswell (2014). The political geographer John Agnew (1987) has identified three aspects of place as location, locale, and sense of place. Location refers to the simple notion of "where"; locale means the material setting within which people conduct their lives; and sense of place addresses the subjective attachment of people to place (Cresswell, 2014). While place has been mostly identified as the binaries such as objective/subjective and material/mental, Edward Soja (1999) have developed the work of French theorist Henri Lefebvre (1991) to challenge these binaries with introducing "thirdspace". Thirdspace as opposed to firstspace (real space) and secondspace (perceived space) is the lived space. This focus on lived world provides a theoretical groundwork for defining place based on lived, practiced, and inhabited space (Cresswell, 2014). Place, therefore, is a structure at a given moment of time, but it is never finished and is always becoming in process, a view that has been informed by structuration school of taught (Cresswell, 2014; Pred, 1984). Similarly, "non-representational theory" approaches place as events and practices rather than interpretations and representation (Thrift, 2008). While we live in pre-structured places, these places are made and remade on a daily basis with the practices within them (Cresswell, 2014). Place has been also seen through the lens of assemblage theory (DeLanda, 2006), where place is a combination of parts that links the inside of a place to the wider world. The relational nature of place can be understood by practices of mobility. It is through our movements that spatial stories are generated (Cresswell & Merriman, 2011) and lived spaces can be identified. With the GPS-enabled mobile devices over the last decade, the virtual world or “cyberspace” has also become place-oriented (Cresswell, 2014). “DigiPlace” has become the “palatial” aspect of the digital world (Zook & Graham, 2007b, 2007a). As a result, different places are created through current information systems that allow us to better understand the real world. Using the information shared through social media, we identify places as 1) un-bounded places of events created through space and time and enriched by text data and 2) bounded entities (census tracts) that are related to other entities trough human flows.

Geographers have mostly identified place as relatively knowable and small scale. Scale has played a significant role in this process and has received a great deal of attention in geography. The definition of scale ranges from the “representative fraction” (McMaster & Sheppard, 2004) to “operational scale” (the areal reach or extensiveness of environmental processes). Scale has been defined at different levels (i.e., global, nation, urban, home, body) (Marston & Smith, 2001; Taylor, 1982) and has been theorized with inclusion of social and ecological processes (Swyngedouw, 2004). The “thingness” of scale and its existence as an ontological structure have been affirmed (Jones, Marston, & Woodward, 2011; Jones, 1998). While there are others who see scale from an epistemological perspective (i.e., Jones et al., 2011; Kaiser & Nikiforova, 2008), our main intension here is to address how using various scalar constructs (Shelton et al., 2014) change the places identified through social media data and ultimately our understanding of the event– situational awareness.

3.2.2 Situational Awareness: Study of Space, Time, Text, and Flows

The literature on the applications of social media data in natural disaster has mainly focused on three areas including event detection, assessment of disaster damage, and situational awareness (Wang & Ye, 2018). Situational awareness (SA) is an important step in reducing human and economic impacts during emergencies (Martínez-Rojas et al., 2018). Situational awareness was originally an aviation term described as “ "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future” (Endsley, 1988 p.97). In basic terms, situational awareness is about understanding what is happening around us. It is information and analysis that can facilitate efficient and effective decision making, which is critical in emergencies (Brady et al., 2013). However, the shortcoming of traditional data in observing human behavior during emergencies has hindered this process (Zou et al., 2018).

With the advent of social media as a major source of real-time data, their proper utilization for enhancing situational awareness has gained more attention. The questions of what, where, and when are of particular importance to disaster decision makers for an informed action during such events. Existing studies have applied spatial clustering techniques such as k-means clustering and kernel density estimation to analyze geospatial information of social media activities during disasters (i.e., Wang et al., 2016). Integrating the geospatial information with content analysis has been more practiced recently for situational awareness, damage assessment, and identifying people at risk (i.e., Wang & Ye, 2018; Kryvasheyeu et al., 2016). However, time has played less role in these studies. The inclusion of time has been mostly limited to the study of disasters before, during, and after, when all social media data in these specific time intervals have been considered and the focus has been on social responses change over time. While spatial proximity in spatial techniques such as clustering plays a significant role, time proximity should also be considered. This is particularly significant when the purpose of using social media during disasters is to provide decision makers with real-time information. Hence, this paper combines spatio-temporal clustering with topic modeling to enhance our understanding of situational awareness during Hurricane Sandy. Additionally, understanding the movement patterns of people during disasters is another essential part of situational awareness. This information is especially crucial for predicting evacuation behavior and preparing the physical infrastructure for that purpose. Therefore, another goal of this paper is to explore the out-flow movements of people– an important topic, yet understudied, during Hurricane Sandy and their association with the underlying social/physical vulnerability of places.

3.3. Methodology

3.3.1 Context of the Study

Hurricane Sandy (Figure 3.1a) started on October 22, 2012 and made landfall near Brigantine, New Jersey, around 8:00 p.m. on October 29, 2012 (NOAA, 2013). Sandy was one of the most destructive hurricanes in the United Stets with 147 direct deaths, around 650,000 damaged or destroyed houses, and approximately 8.5 million customers without power during and after the storm (NOAA, 2013) .The record levels of storm surges were observed in the states of New York, New Jersey and Connecticut, especially in and around New York City (NOAA, 2013).

3.3.2 Twitter Data Collection and Processing

4.4 million Twitter messages from affected coastal counties of New York, New Jersey, and Connecticut were collected from Oct. 22 to Nov. 7, 2012 using Firehose streaming API via GNIP (Gnip APIs). Tweets generated within NYC (582161) were selected for this study. Following the work of Tsou, Zhang, & Jung (2017), we manually reviewed the “generator” field in the dataset to identify the bots or cyborgs with commercial purposes resulting in removal of

11922 tweets. These noises were advertisement, traffic, news, weather, and job tweets including dlvrt.it, pinprick on iOS, Squarespace, COS App, kickalert, Beer Menus, dine here, Dance Deets, 511NY-Tweets, TTN NYC traffic, TweetMyJOBS, and SafeTweet by TweetMyJOBS (Pourebrahim, Sultana, Niakanlahiji, & Thill, 2019). For the purpose of spatial clustering and topic modeling, only tweets in the evacuation zones of NYC (Figure 3.1b) were considered resulting in a total of 160565 usable tweets. The movement analysis was conducted at census tract level for all tweets generated within NYC during the hurricane period (10/29/2012 – 10/31/2012).

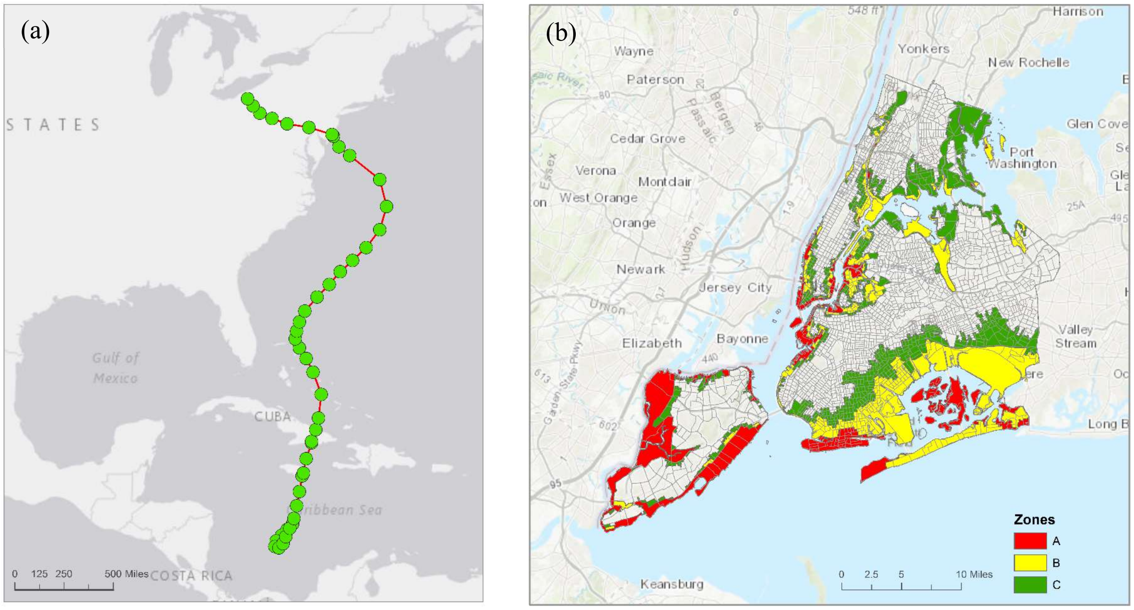


Figure 3. 1. Hurricane Sandy Path (a) and NYC Evacuation Zones (b).

3.3.3 Spatio-Temporal Clustering

Density-based spatial clustering with noise (DBSCAN) is an unsupervised data clustering algorithm, which does not require a prior knowledge of the number of clusters (Ester, Kriegel, Sander, & Xu, 1996). DBSCAN is useful for discovery of clusters with arbitrary shape and can remove the noise outliers from the clusters making it a useful technique in analyzing the complex twitter patterns (Huang, Li, & Shan, 2018). DBSCAN, however, only accounts for spatial distance. Therefore, spatio-temporal clustering ST-DBSCAN (Birant & Kut, 2007) was used so that temporal distance of data can be also taken into consideration. ST-DBSCAN was applied to cluster the tweets of each day. To identify clusters of Twitter activities at fine spatial granularity, we defined a search window of 300 meters, almost equal to average length of census blocks in NYC. The time value was set to 60 minutes so that we can identify the clusters in hourly intervals. Although some events last long, many cannot maintain that many points in every 60-min period (Huang et al., 2018). The minimum number of tweets to form a cluster was selected to be 10 to avoid non-significant, random chats among Twitter users (Huang et al., 2018) and to ensure that sufficient number of people are present in each cluster so that the topic of discussions can be easily identified. After applying ST-DBSCAN, two other parameters were used to filter the clusters. First, minimum number of users in each cluster was set to 2 and if more than half of the tweets in a cluster were from one user that cluster was removed. This filtering helped us to avoid non-significant clusters where a single user generates most of the tweets.

3.3.4 Topic Modeling

To identify if a specific event was happening in a place, topic modeling was applied to the tweets' texts within each identified cluster of activities. Latent Dirichlet Allocation (LDA) is a generative probabilistic model of a corpus, where documents are represented as random mixtures over latent topics and each topic is characterized by a distribution over words (Blei, Ng, & Jordan, 2003). Each document in LDA is assigned a mixture of topics.

$$P(Z/W, D) = \frac{W_{Z+\beta W}}{total\ tokens\ in\ Z+\ \beta} {*D}_{Z+\alpha}$$

For each topic *Z*, the probability *P* that word *W* came from topic *Z*, is calculated by multiplying the normalized frequency of *W* in *Z* with the number of other words in document *D* that already belong to *Z* (Sit, Koylu, & Demir, 2019). *β* and *βw* are hyper parameters that are used to incorporate the probability of word *W* belonging to topic *Z*. LDA goes through each word in documents iteratively and reassign each word to a topic until the model reaches an equilibrium (Sit, Koylu, & Demir, 2019; Blei, Ng, & Jordan, 2003). The LDA topics in our case study refer to the aspects of a specific event in clusters and each tweet is a document. Before applying LDA, tweets were cleaned to remove URLs, stop words, punctuations, extraneous words and so on. We also discarded the words that were appeared in less than 1% or more than 90% of tweets to avoid words that could not give much information. Grid search was applied to identify the best topic model parameters (number of topics and learning rate) and then topics were visualized.

3.3.5 Movement Analysis

We extracted Twitter flows by converting the Twitter dataset into an origin-destination matrix. Origin represents a census tract (A) where an individual is located at time t and destination is the census tract (B) where the same individual is located at time t+1. The direction of flow was considered from census tract A to census tract B. The individual flows were aggregated for each census tract to identify the total number of out-flows (flows leaving the census tract) for each census tract. These flows were then entered to an ordinary least squares regression model (OLS) as the dependent variable with social and physical vulnerability as independent variables. Due to presence of non-stationarity, we applied geographically weighted regression (GWR) (Lu, Charlton, Harris, & Fotheringham, 2014) for taking into consideration the spatial dimension of data. While the purpose of OLS and GWR models are explanatory analysis, machine learning algorithms provide more inferential power. Random forest (RF) algorithm (Breiman, 2001) is among the most efficient ensemble methods for regression problems (Ghasri, Hossein Rashidi, & Waller, 2017). Random forest is an ensemble of decision trees that grow in parallel on a subsample of the training data. RF is an ensemble of C trees $T_1(X)$, $T_2(X)$, . . .,

$T_C(X)$, where $X = x_1, x_2, . . ., x_m$ is a m-dimension vector of inputs and the outputs are $\hat{Y}_1 = T_1(X)$, $\hat{Y}_2 = T_2(X), . . ., \hat{Y}_C = T_C (X)$. The average of all trees' outputs ($\hat{Y}_1$ to $\hat{Y}_C$) will be the final prediction $\hat{Y}$ (for more details see Pourebrahim et al., 2019). An RF model is built from 2/3 of the data (in-bag) and excluded data (out-of-bag) are used for identifying the prediction error (Ostmann & Martínez Arbizu, 2018a). The out-of-bag (OOB) data also eliminates the need for a separate test set and is used to identify the importance of predictors (Breiman, 2001). RF, however, is still "aspatial" in nature. Georganos et al. (2019) developed a geographical random forest (GRF) by integrating the RF algorithm with the geographically weighted regression (GWR). In comparison to GWR, GRF allows for modeling non-stationarity coupled with a flexible non-linear model that is hardly overfit due to bootstrapping and relaxes the assumptions of traditional Gaussian statistics (Kalogirou & Georganos, 2019) . The difference between RF and GRF can be described by a simplistic version of a regression equation (Georganos et al., 2019) :

$$Y_i = ax_i+e, i = 1:n$$

Where Y_i is the dependent variable value for the ith observation, ax_i is the non-linear prediction of RF based on independent variables (x), and e is the error term. In GRF, the equation can extend to:

$$Y_i = a(u_{i,,}\ \ v_{i,}\ )x_i+e, i = 1:n$$

Where $a(u_i, v_i)$ is the prediction of an RF model that is calibrated on location i with coordinates of (u_i, v_i). A local RF model is calculated in each data point by defining a neighborhood area (kernel). Finally, the model provides the global and local goodness of fit (for more details see Georganos et al., 2019). We compared the results of OLS, GWR, global RF, and local RF (GRF) to identify the best model. Root mean squared error (RMSE) and R-squared (R^2) are reported. The OOB results are reported for the RF and GRF models.

To identify social vulnerability of a census tract, the 14 variables of the Center for Disease Control and Prevention's (CDC) Social Vulnerability Index (SVI) were used (Table 1) (Flanagan, Gregory, Hallisey, Heitgerd, & Lewis, 2011). The CDC's SVI shows areas where certain social conditions may affect the ability of a community to prevent consequences of disasters (Wang, Lam, Obradovich, & Ye, 2019; Flanagan et al., 2011). We applied principal component analysis (PCA) due to high correlation among the variables of the CDC's SVI to reduce the dimension of data. PCA is a multivariate technique that analyzes a set of corelated variables and display it as a set of new orthogonal variables named principal components (Abdi & Williams, 2010). The scores for the component showing the highest total variance were generated for all census tracts. These scores were used as independent variable in the final model with higher score showing higher vulnerability. Additionally, Euclidean distance from census tracts centroids to nearest shoreline was calculated as physical vulnerability (Wang, Lam, Obradovich, & Ye, 2019).

Table 3. 1 Social Vulnerability Variables (Source: CDC's Social Vulnerability Index)

Socioeconomic Variables	Proportion of persons below poverty estimate
	Proportion of civilian (age 16+) unemployed estimate
	Per capita income estimate, 2006-2010 ACS
	Proportion of persons with no high school diploma (age 25+) estimate
Household Composition Variables	Proportion of persons aged 65 and older
	Proportion of persons aged 17 and younger
	Proportion of single parent households with children under 18
Minority Status/Language Variables	Proportion minority (all persons except white, non-Hispanic)
	Proportion of persons (age 5+) who speak English "less than well" estimate
Housing/Transportation Variables	Proportion of housing in structures with 10 or more units estimate
	Proportion of mobile homes estimate
	Proportion of households with more people than rooms estimate
	Proportion of households with no vehicle available estimate
	Proportion of persons in institutionalized group quarters

3.4 Results and Discussions

To identify places as experienced by people during the Hurricane Sandy, we first applied a spatio-temporal clustering analysis, where both spatial and temporal distance were taken into consideration. the spatio-temporal information allowed us to identify where and when specific events are happening during the hurricane. The combination of this information with text analysis enabled us to determine the particular event at place. Here, the clusters for two days of October 29 and October 30 are presented. Most of clusters were identified in the Manhattan during these days because of high volume of tweet activities in the area. The larger clusters show the larger number of tweet activities. These clusters also last for a longer period and usually are representative of specific events that caught people's attention more. For instance, the two large clusters 1 and 2 in lower Manhattan on October 29 (Figure 3.2), are related to Battery Park, Financial District, and West Village flooding. People were also talking about the rain, water height, power, and 14th street flooding in the East Village on October 29 (cluster 3).

Three topics were identified in each cluster as the result of grid search. The intertopic distance map (via multidimensional scaling) and the top 30 most relevant words for each topic are presented (for more details see Sievert & Shirley, 2014) for specific events in October 29 and 30. Figure 3.3 shows the topic identified in cluster 1 of October 29. In the first topic, the word “York”, “battery”, “park”, “financial”, and “district” are the top 5 frequent words. The word “flooding” is also among the top 30 words. The words “street”, “wall”, “lower”, “manhattan”, “bilding”, “power”, “flooded” are observed in the second topic. In October 29, hurricane Sandy flooded the Financial District (Figure 3.4) resulting in the closing of stock exchanges. Battery Park and Wall Street Plaza in the Financial District were also closed because of flooding.

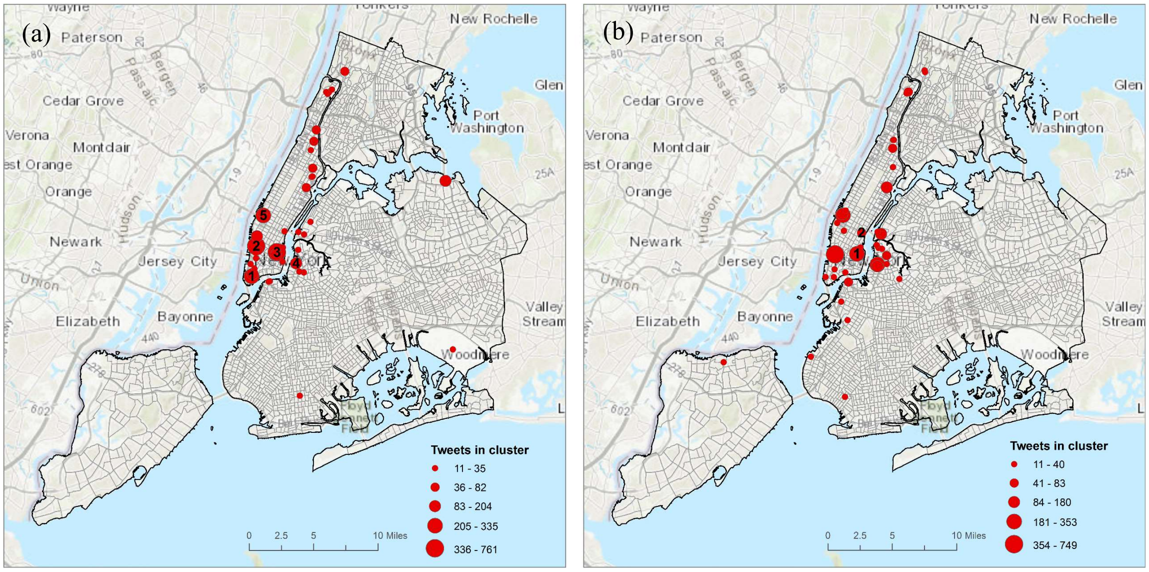


Figure 3. 2. Cluster of Activities on October 29 (a) and October 30 (b) (300-Meter Scale).

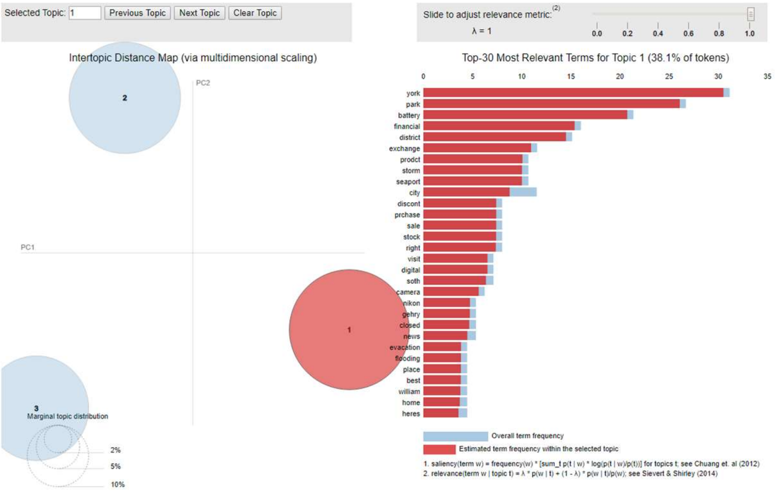

Selected Topic: 1
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
2
1
3
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 1 (38.1% of tokens)
0
5
10
15
20
25
30
35
york
park
battery
financial
district
exchange
prodct
storm
seaport
city
discont
prchase
sale
stock
right
visit
digital
soth
camera
nikon
gehry
closed
news
evacation
flooding
place
best
william
home
heres
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 3. 3. Cluster 1, October 29.

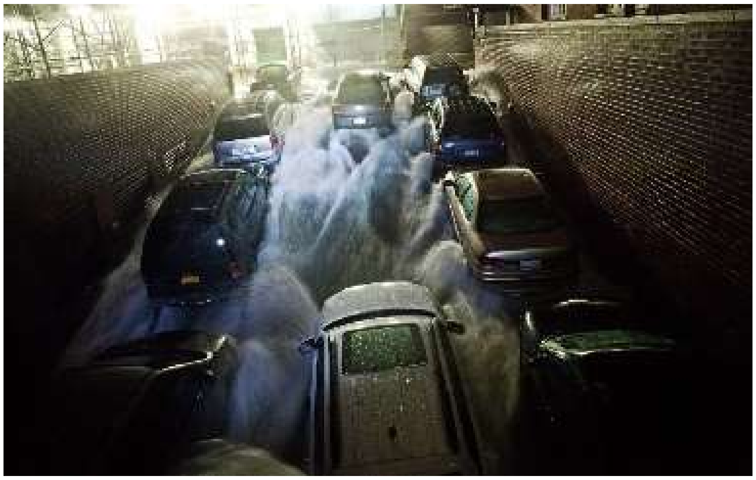


Figure 3. 4. A Parking Garage in the Financial District, October 29 (Source: BBC News).

Several piers in the Hudson River Park experienced extensive damage with the hurricane hit, infrastructures were destroyed resulting in almost $10 million in damage, and park was closed. There was also flooding and damage in the West Village, Meatpacking District, and Chelsea, a block from Hudson River. Our analysis identified these places in the most relevant words of cluster 2 (Figures 3.2a & 3.5). People were reporting the wind, rain, and power outage in these places. The words “Chelsea”, and “bilding” in the third identified topic of this cluster refers to a building facade collapse at Chelsea (Figure 3.6). Similarly, Brooklyn’s Williamsburg experienced high winds and ferry piers on the East River were closed (Cluster 4, Figure 3.2a). A construction site in the area felled due to the high winds. The identified topics in Figure 3.7 are representative of these events. Another important event in Manhattan was a crane collapse at 57th street (Figure 3.8). The third topic (Figure 3.9) identified in the cluster 5 of October 29 had the word “crane” as the top most relevant term. Although the location of the collapse is not identified in the top words, the words “crane”, “57th”, “collapsed” were among the 30 relevant terms of the topics identified for another cluster located around Brooklyn Bridge. This cluster was shaped at the same temporal proximity and addressed the exact location of the crane collapse.

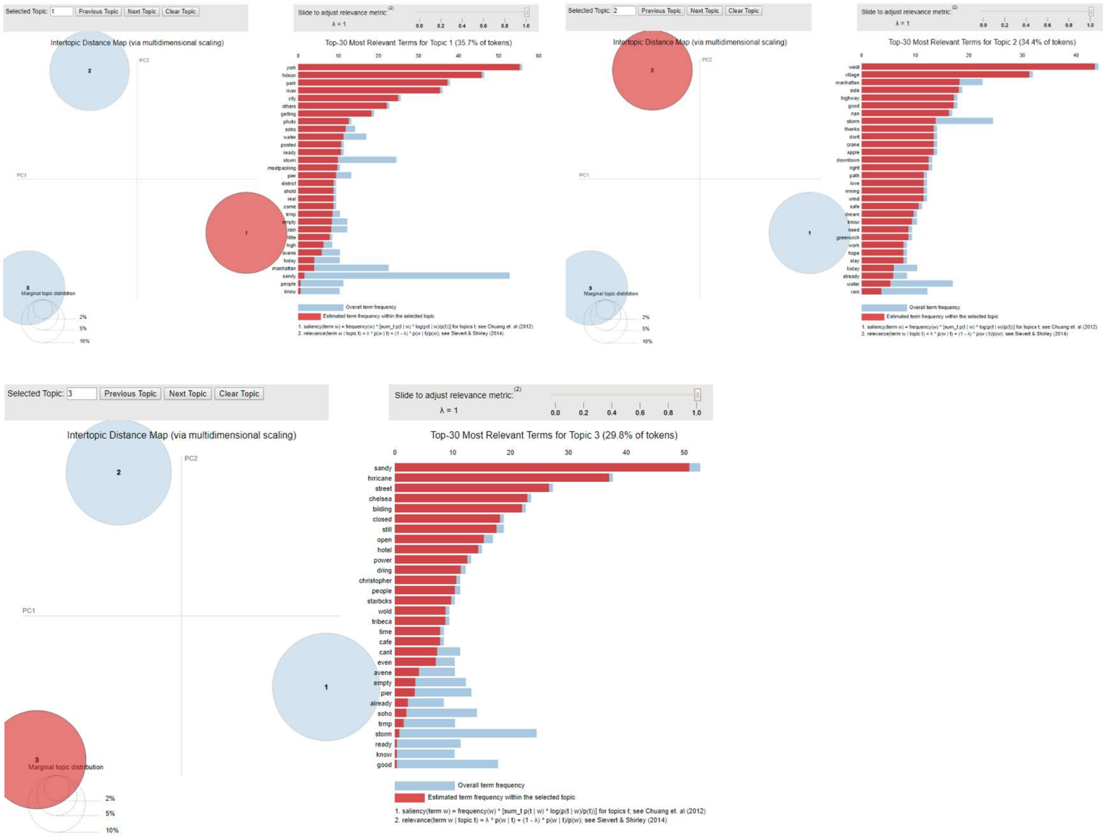


Figure 3. 5. Cluster 2, October 29.

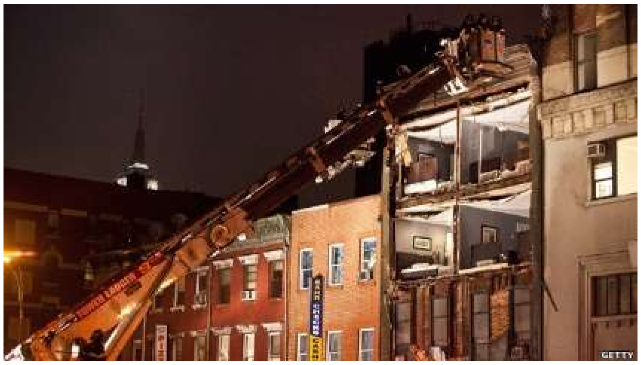
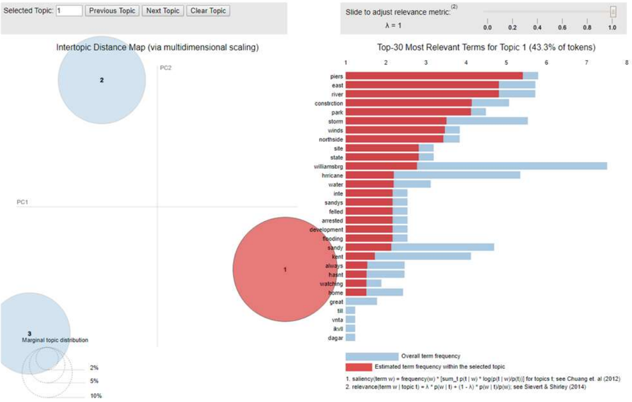


Figure 3. 6. Building Façade Collapse, Chelsea, October 29. (Source: BBC News).



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 3. 7. Cluster 4, October 29.

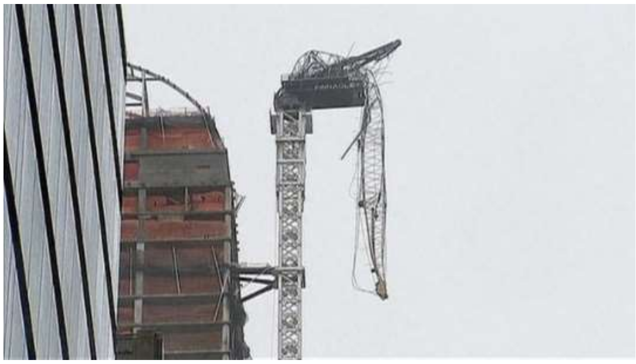
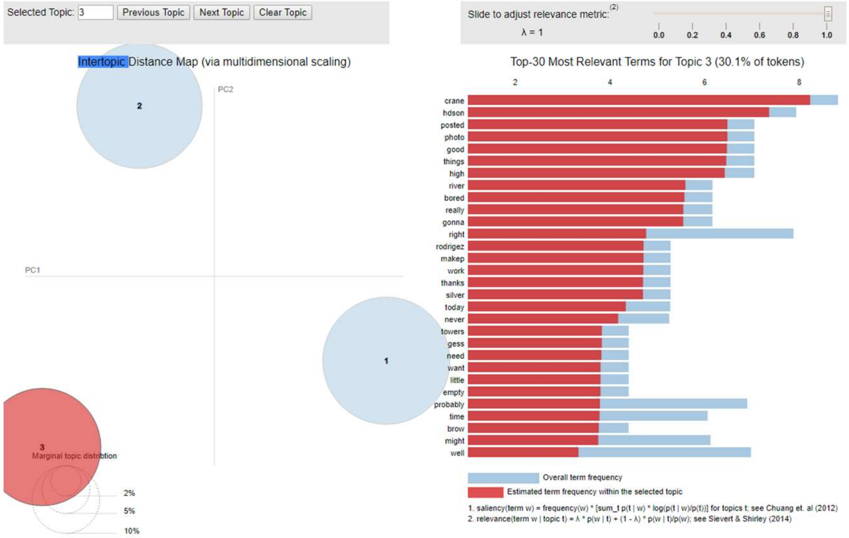


Figure 3. 8. Crane Collapse at 57th Street, October 29. (Source: BBC News).



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 3. 9. Cluster 5, October 29.

14th street in East village is another place identified in the analysis. The street experienced flooding on October 29 (cluster 3, Figure 3.2a) and power outage on October 30 due to an explosion in a substation of Consolidated Edison utility provider (Figure 3.10). The words “power”, “east”, “village”, “street”, “14th”, and “explosion” are among the top relevant terms (Figure 3.11) in the cluster 1 of October 30 (Figures 3.2b). The power outage of the East 33rd street and subway flooding (Figure 3.12) in the cluster 2 (Figure 3.2b) are other examples of identified events in October 30.

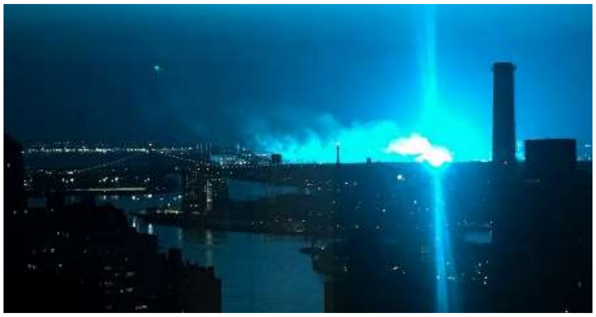
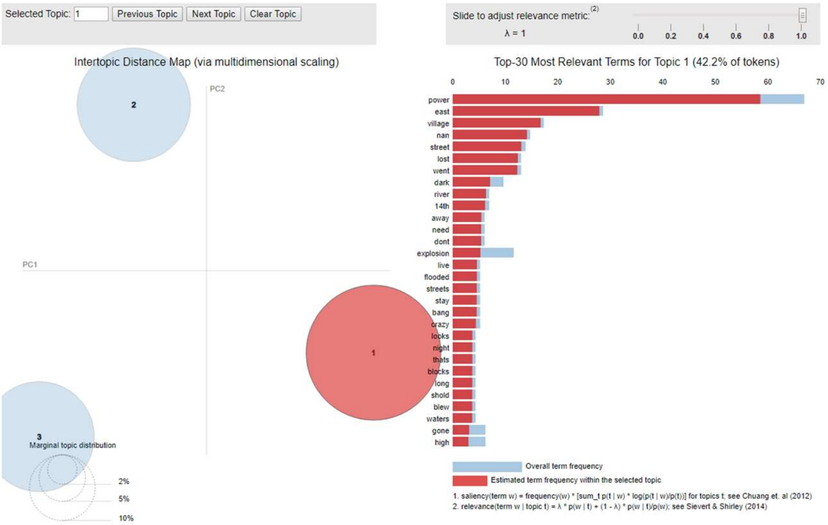
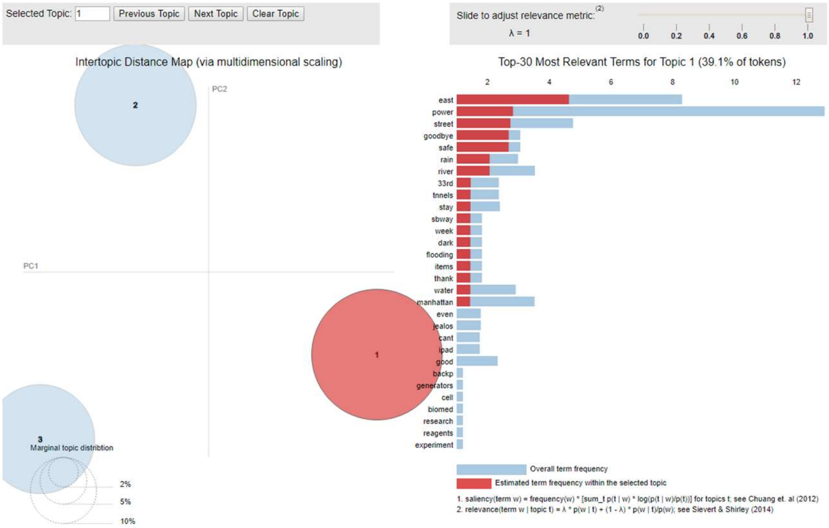


Figure 3. 10. Explosion in Consolidated Edison Utility Provider Substation, October 30. (Source: HexByte).



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 3. 11. Cluster 1, October 30.



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 3. 12. Cluster 2, October 30.

Since the majority of clusters were identified in Manhattan and north of Brooklyn, we changed the spatial resolution from 300 meters to 600 meters to explore if other clusters can be recognized (Figure 3.13). The coarse spatial granularity allowed us to identify some new clusters in Brooklyn and Staten Island, where tweet activities were lower. Red Hook, for example, was one of the most affected communities during the Hurricane Sandy. Red Hook residents experienced power, heat and running water outages for several weeks. While no events related to Red Hook was identified in the previous analysis, a cluster of topics (cluster 1, Figure 3.13a) related to this event was identified in the coarse granularity analysis of October 29 (Figure 3.14). Another example is cluster 2 of October 29 (Figure 3.13a) around Gowanus canal with the identified words of “canal”, “gowans”, “bridge”, “carroll”, “street”, “emergency”, “flooded” (Figure 3.15). During the hurricane, Gowanus Canal broke its bank and flooded over many streets. The canal was among the most contaminated water bodies due to industrial activities. There were also other events identified through the coarse granularity analysis of October 30 (Figure 3.13b). Example are power outage and flooding in clusters 1, car fire in cluster 2, and uprooted trees at Queens’ Astoria Park (Figure 3.16) in cluster 3 (Figure 3.13b). The identified topics of cluster 3 are presented in Figure 3.17.

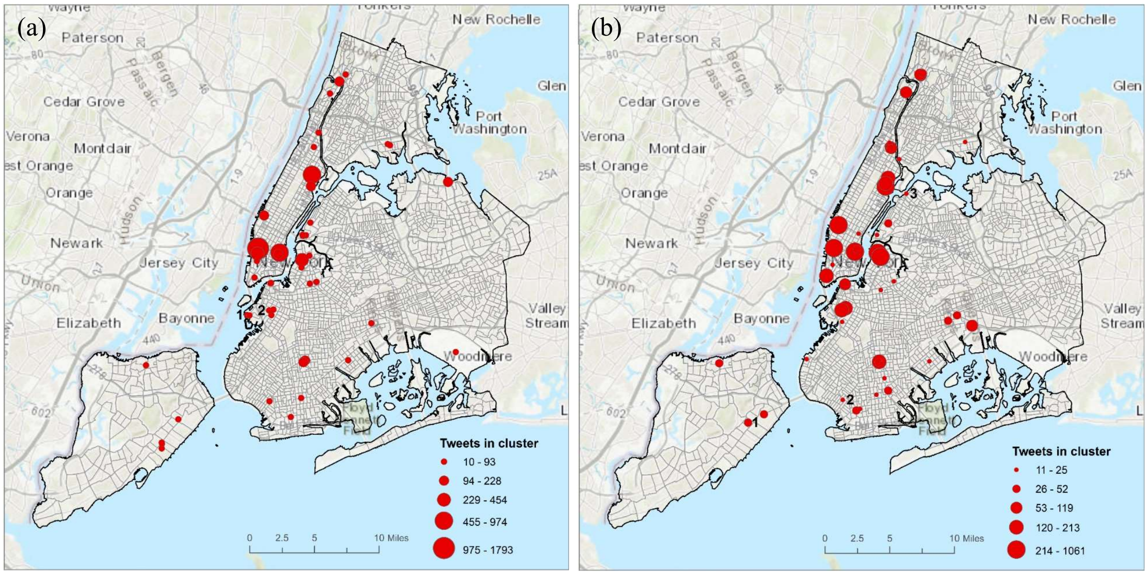
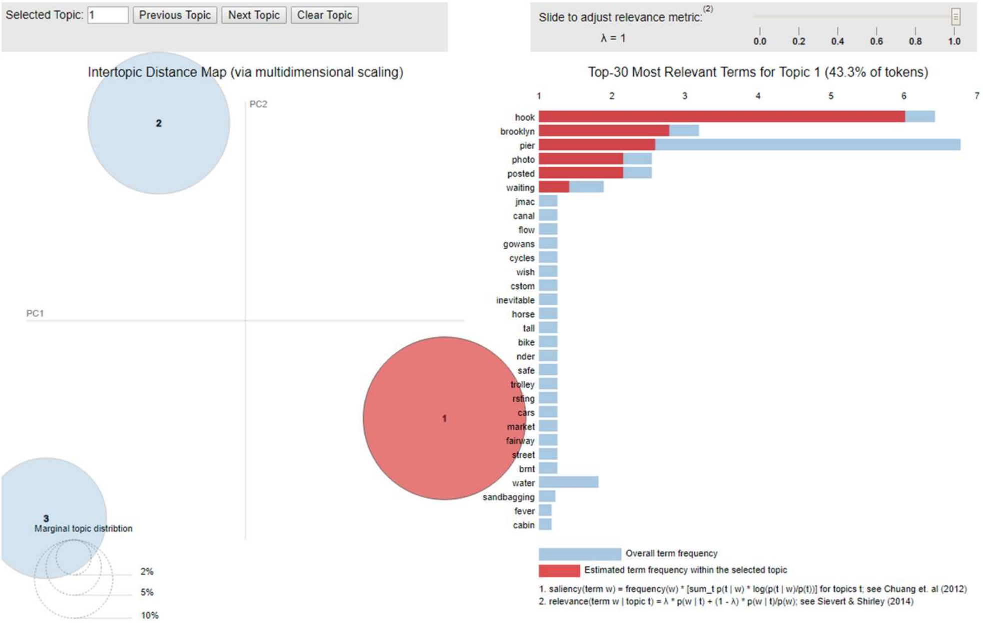
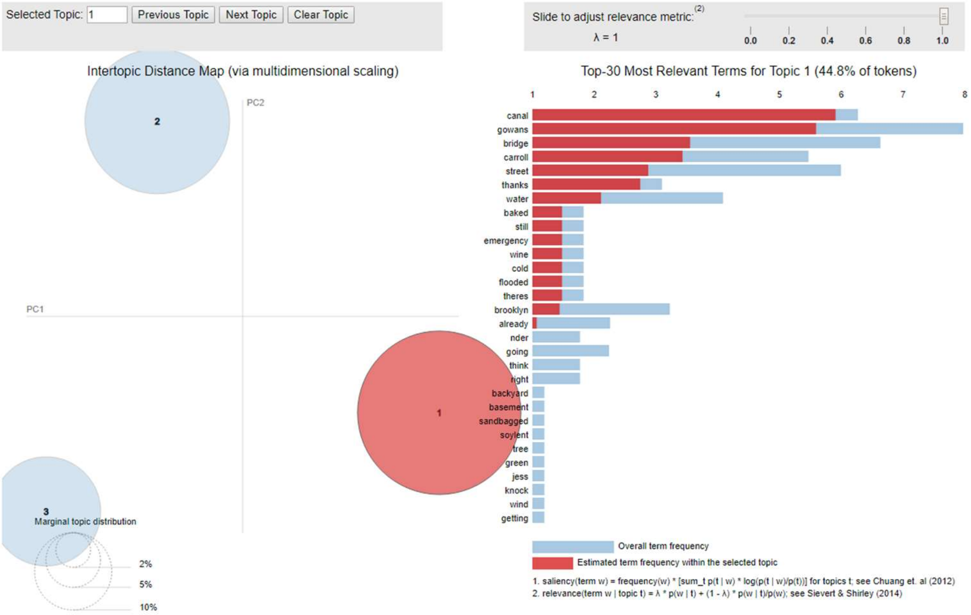


Figure 3. 13. Cluster of Activities on October 29 (a) and October 30 (b) (600-Meter Scale).



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 3. 14. Cluster 1, October 29 (600-Meter Scale).



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 3. 15. Cluster 2, October 29 (600-Meter Scale).

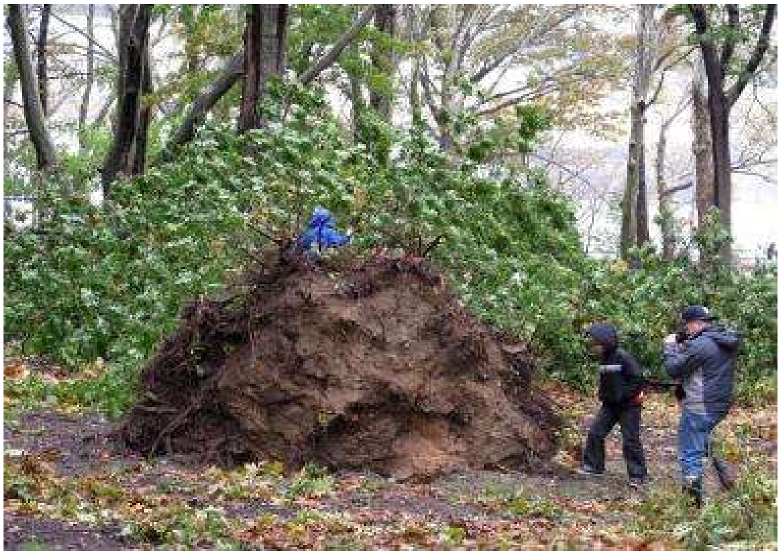
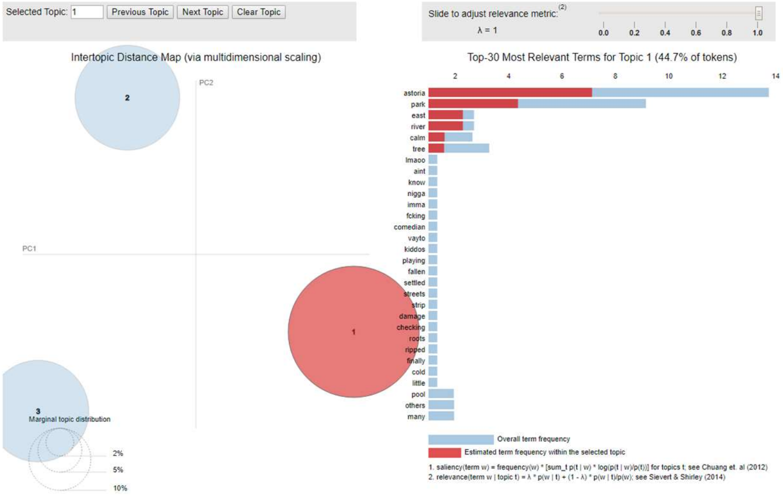


Figure 3. 16. Uprooted Tree in Astoria Park, October 30. (Source: DNAinfo).



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 3. 17.Cluster 3, October 30 (600-Meter Scale).

Since our analysis was conducted at fine spatial and temporal scale, the examples here highlight the potential of developing automated models during natural disasters, so that an enhanced situational awareness can be achieved at near real time. These examples also demonstrate the importance of social media analysis with a place-based theoretical approach.

Although it is useful to identify the material phenomena of activities in Cartesian space by such data, stopping there neglects the way that these data are reflective of people's experience of places (Shelton et al., 2014). With a focus on the qualitative text data in this paper, the potential of identifying the experience of place embodied in the data was identified. We were able to go beyond the simple mapping of terms and activities and identify the meaning embodied in the data and places that can be inferred. The identified places were the result of either people experience in an exact location (i.e., specific street) or people talking about another place. For example, while the topic of "crane" was identified in the cluster 5 of October 29, the exact location of the event (57th street) was actually mentioned in another cluster located around Brooklyn Bridge. There were also similar examples, where people in a cluster located in the East 20th street were talking about the events happening in the East River Park. These examples show how places can be identified through their relationship with other places. These relational places also acknowledge the potential for people experiences to be spatially distanciated from the exact locations of specific evets (Shelton et al., 2014).

Although change of scale in areas with lower tweet activities helped us to identify new clusters, generally clusters were present at specific areas, which lead us to the data completeness problem. Many southern parts of NYC that were close to the coastline had significant damages during Hurricane Sandy. However, our analysis could not identify those places because of lower Twitter activities. Same problem can be observed in the analysis of movement. Figure 3.18a shows the out-flows in the NYC during Hurricane Sandy period. The higher flows are concentrated in Manhattan, reflecting the spatial nature of these data. This can also be inferred from the low predictive power of the OLS model and the statistical significance of JB and Omni tests showing the non-normality of residuals (Table 2). Moran's I tests also showed the spatial autocorrelation of residuals. To account for spatial non-stationarity of data, GWR was conducted.

The results show an improvement in the model. R^2 increased by 52% and RMSE decreased by 6.44. the GWR results were also better than the RF and GRF model as presented in Table 3. It should be noted that GRF model also had better results compared to RF, reinstating the better performance of spatial models compared to non-spatial ones.

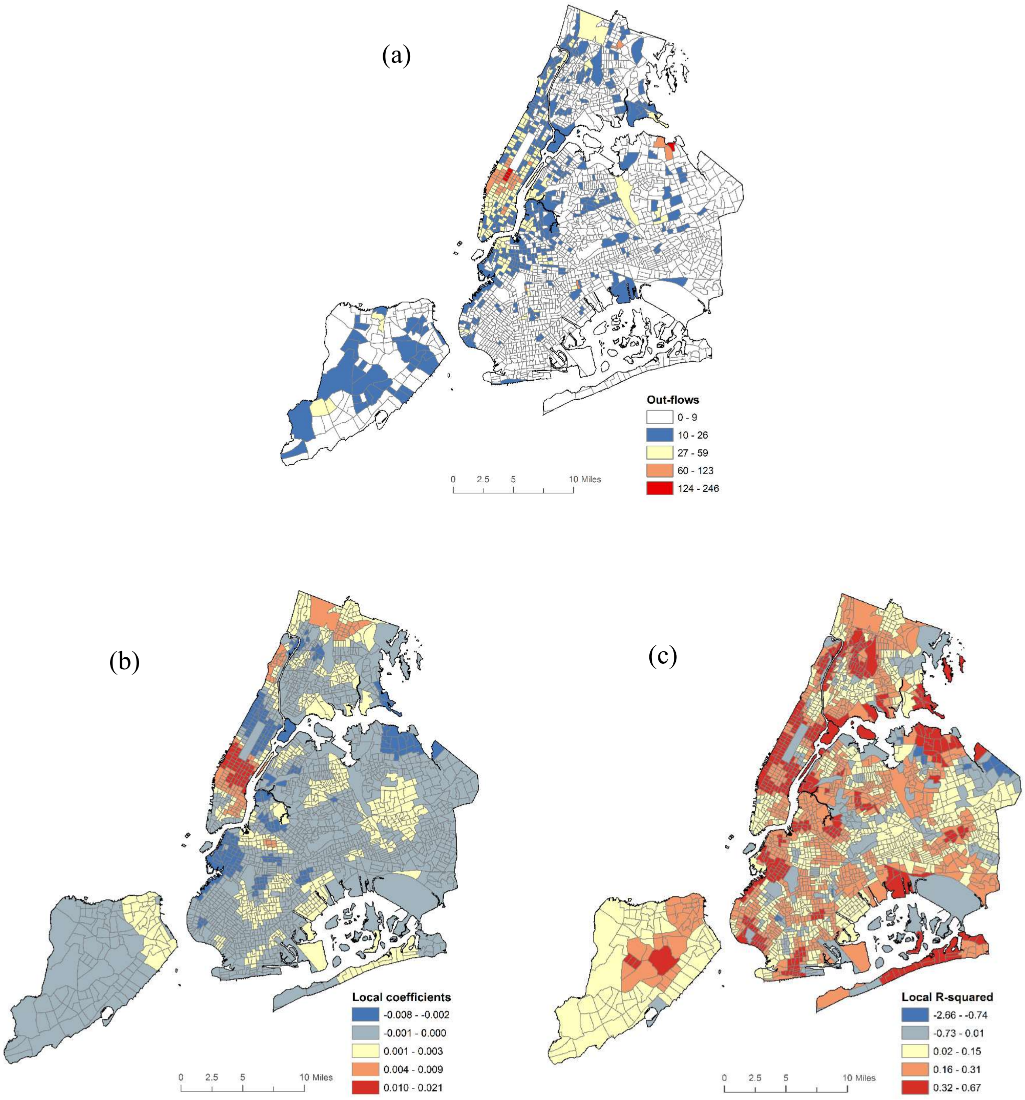


Figure 3. 18. Census Tracts' Out-Flows (a), Local Coefficients of Distance to Shoreline (b), and Local R^2 Results (c).

Table 3. 2 OLS Results.

R-squared	0.110	**Adj. R-squared**	0.110
F-statistic	131.1	**p-value (F-statistic)**	2.16e-54
Jarque-Bera (JB)	214301.304	**p-value (JB)**	0.00
Omnibus	2265.192	**p-value (Omnibus)**	0.000
Moran's I	0.41	**p-value**	0.000
	coefficient	**standard error**	**p-value**
Intercept	14.0596	0.540	0.000
Physical Vulnerability	-0.0008	7.26e-05	0.000
Social Vulnerability	-4.2041	0.339	0.000

Table 3. 3 Performance of Models.

Model	**R^2**	**RMSE**
OLS	0.11	15.55
GWR	0.63	9.11
RF (OOB)	0.16	15.15
GRF (OOB)	0.49	11.81

The GWR results demonstrate that while both social and physical vulnerability were important in the models, social vulnerability had a global negative effect. Census tracts with higher social vulnerability had lower rate of out-flows. Distance to shoreline, however, had a local impact with higher out-flows in some places close to shoreline (i.e., east of Central Park, east of Queens, and west of Brooklyn) and lower out-flows in some other places close to shoreline (i.e., West of Central Park) (Figure 3.18b). The local R^2 also varies from the places, where regression does not fit the data to places with higher prediction power. Census tracts with higher R^2 are scattered within NYC with a higher concentration in Manhattan, east and south of Queens, and different places in Brooklyn (Figure 3.18c). The spatial analysis of human flows showed the creation of places from a different angle. These places that are shaped through their relationship with each other play different roles during the hurricane, with some having more out-flows and others having less. The extent to which the out-flows in these places can be predicted through physical vulnerability of the place also varies significantly.

3.5 Conclusions

The need for theoretical work to reconcile the world of space (traditional GIS) and the world of place (social media) has been identified in the geographic literature (Yang et al., 2016). Our research has insights for both theoretical and computational approaches to social media analysis. With a focus on the qualitative experience of place embodied in the Twitter data coupled with concepts of space, place, and scale, we examined social media analysis of disasters in the GIS domain. We first identified places as unbounded clusters of events created through space and time and enriched by text data. These places were formed by people's experience of an event in an exact location (i.e., specific street). There were also places identified through their relationship with other places, when people in a spatial cluster were talking about specific event in another cluster. In another definition of place, we identified place as bounded entities (census tracts) that are related to other entities trough human flows. These relational places were identified through people's movements, where places of high and low out-flows were observed. The identified places in this study represent the "locale" or "thirdspace" perspective of place, where place is defined through lived, practiced, and inhabited space, and it is a structure at a given moment of time that is always becoming.

Identifying events during natural disaster is crucial for a better response in real time. Developing automated models to enhance situational awareness, therefore, can be a future direction. In addition, the movement patterns of the population during Hurricane Sandy were analyzed in this study. While the better performance of spatial models was identified, predicting movement during natural disasters is a difficult task as the variables contributing to the movements are different from non-chaotic environment. Moreover, unavailability of data to validate the results makes the predictions unreliable. Combining social media data with other sensors’ data (i.e., remote sensing) to construct concrete plans and strategies for strengthening resilience (Zou et al., 2018) remains for future research.

CHAPTER IV

TRIP DISTRIBUTION MODELING WITH TWITTER DATA[3]

4. 1. Introduction

Worldwide increases in traffic congestion and air pollution in urban areas presents a need to better understand mobility patterns of urban populations and their travel demands (Shirzadi Babakan, Alimohammadi, & Taleai, 2015; Y. Yang, 2013). A number of studies have examined either individual or collective mobility patterns at different spatial scales (i.e., Beiró, Panisson, Tizzoni, & Cattuto, 2016; Hawelka et al., 2014; González, Hidalgo, & Barabási, 2008). Mobility information of individuals can be aggregated to study the frequency of travel between different regions, as represented by origin-destination (OD) matrices (Barbosa et al., 2018). An OD matrix provides population flow patterns (trip distribution) studied for diverse purposes such as traffic forecasting, resource allocation, prediction of migration flows, and epidemic spreading (Beiró et al., 2016). Therefore, improving flow estimations has become critical across various domains of application (Barbosa et al., 2018).

Various trip distribution models have been developed over past decades to estimate population flows with greater accuracy (Simini, González, Maritan, & Barabási, 2012; de Dios Ortuzar & Willumsen, 2011; Roy & Thill, 2003; Wilson, 1998; Wilson, 1970; Zipf, 1946). Most of these models are heavily dependent on conventional data such as population censuses or travel diary surveys. The emergence of social media and location-based services in recent years has

[3] Pourebrahim, N., Sultana, S., Niakanlahiji, A., & Thill, J.-C. (2019). Trip distribution modeling with Twitter data. *Computers, Environment and Urban Systems*, *77*(June 2019), 101354. https://doi.org/10.1016/j.compenvurbsys.2019.101354

introduced new opportunities to the field of transportation (Yang, Jin, Cheng, Zhang, & Ran, 2015). Geospatial big data such as taxi trajectories, mobile phone records, and social media messages have attracted scholars to observe, understand, and visualize (i.e., Karduni et al., 2017) human activities in cities at fine spatio-temporal scales (Liu et al., 2015). These data significantly improve the visualization of human mobility patterns, yet there is a need to better understand and contextualize them in different steps of the travel demand modeling framework (Anda, Erath, & Fourie, 2017).

Traditionally, the gravity model and its derivatives have been used as the most reliable approach to predict trip distribution at fine spatial scales, such as commuting flows within cities (Lenormand, Bassolas, & Ramasco, 2016). The potential for developing hybrid approaches that integrate the vast volume of social media data with the gravity model has recently been noted (Beiró et al., 2016). While traditional models have used statistical methods rooted in sound mathematical foundations, they have been unable to account for nonlinearities and other irregularities in data (Golshani, Shabanpour, Mahmoudifard, Derrible, & Mohammadian, 2018). To alleviate these and other issues, Machine Learning (ML) techniques have been applied in different urban and transportation domains (i.e., Karimi et al., 2019; Ghasri et al., 2017). A significant body of literature exists at this time where various ML techniques have been evaluated on their ability to model travel demand, such as artificial neural networks (ANNs) (i.e., (Pouebrahim, Sultana, Thill, & Mohanty, 2018; Ding, Wang, Wang, & Baumann, 2013; Tillema, van Zuilekom, & van Maarseveen, 2006; Mozolin, Thill, & Lynn Usery, 2000) and tree-based ensemble methods (i.e., Ghasri et al., 2017; Rasouli & Timmermans, 2014). While random forests (RFs) (Breiman, 2001) have been identified among the most advanced and most efficient ensemble methods for data classification and regression (Ghasri et al., 2017), they have so far been used in only a few studies of travel demand. Ghasri et al. (2017) and Rasouli and

Timmermans (2014) have reported promising results with RF modeling of trip generation and modal split, yet their suitability and usefulness in trip distribution analysis remains to be thoroughly assessed.

Given the current state of research, our objective is to compare the performance of gravity, neural network, and random forest models of commuting trip distribution while combining both traditional and social media data. We also evaluate how information on personal mobility derived from social media affects commuting trip distribution by identifying the importance of different variables. To the best of our knowledge, this paper is one of the first to use machine learning approaches in trip distribution forecasting with social media data. The main contributions of this paper are threefold: (1) revealing the potential of social media data in trip distribution modeling at census tract level; (2) using machine learning techniques to predict trip distribution at census tract level; and (3) comparing the performance of gravity, neural network and random forest models to identify the best model for predicting trip distribution at census tract level. The paper is organized as follows. The review of related work is provided in section 4.2, followed by a presentation of the study area, data sources and methodology in section 4.3. Results are presented in section 4.4, with a discussion and concluding remarks in sections 4.5 and 4.6.

4.2. Related Work

4.2.1 Travel Demand Modeling

The relationship between personal mobility flows and a range of personal and environmental factors has been studied to determine future travel demands within cities (Barbosa et al., 2018). Travel demand modeling has long been dominated by the four-step model with its steps being trip generation, trip distribution, modal split, and traffic assignment (Mcnally, 2007). The objective of this model is to estimate the traffic in the transportation networks. The model first identifies the amount of travel produced by each traffic zone in the area (trip generation), which is then distributed among other zones (trip distribution). Trip distribution accounts for network effects on personal mobilities and responds to the connectivity qualities of travel opportunities across the urban space. A broad family of models of spatial interaction, including the ubiquitous gravity model, has been developed to analyze and forecast urban trip distribution (i.e., Simini et al., 2012; de Dios Ortuzar & Willumsen, 2011; Roy & Thill, 2003; Wilson, 1998; Wilson, 1970; Zipf, 1946).

In his pioneering work, Zipf (1946) suggests that the number of persons moving between two communities is proportional to the product of their populations and is inversely related to the distance between them. This basic model has been extended over time by adding factors other than population. These factors that determine trip production rate at origins and trip attraction volumes at destinations have been examined in past studies. Population, housing type, household income, and car ownership have been identified as important factors determining volumes of trip production in the origin zone for commuting trips (Berger, 2012). Employment, density and land use, and points of interest have been identified as attraction factors of the destination zones (Yang, Herrera, Eagle, & González, 2015; Berger, 2012).

4.2.2 Social Media and Human Mobility

Cities with large populations are early adopters of new information and communication technologies (Barbosa et al., 2018). As such, they have more mobile phone users, generating a huge volume of data through various social networking applications. Social media data, particularly Twitter, have attracted many scholars to explore human mobility at various spatial scales. For example, Hawelka et al. (2014) employed Twitter data to examine mobility patterns of international travelers. Several measures were explored including mobility rate, radius of gyration, diversity of destinations and the balance of inflows and outflows. Similarly, Kurkcu, Ozbay, & Morgul (2016) studied radius of gyration and user displacement for Twitter users in

New York City. The research showed that mobility patterns of Twitter users follow the Lévy Flight and that the mobility flows estimated from Twitter posts are similar to ground-truth home-to-work trips at the county level. In their analysis of human mobility and activity patterns, Hasan, Zhan, & Ukkusuri (2013) studied the distribution of different activity categories using Twitter data, as well as the temporal, spatial and frequency distributions of places visited in New York, Chicago and Los Angeles. Liu, Zhao, Khan, Cameron, & Jurdak (2015) used Twitter data to investigate population distribution and human mobility flows in Australia at national, state, and metropolitan scales. Their study found that population distribution can be estimated from Twitter data at coarse spatial granularity.

While mobility patterns identified from Twitter data have been commonly reported through simple measures such as those mentioned above, a few studies have used more advanced modeling perspectives. Liu et al. (2015) reported a high correlation between the mobility flows extracted from Twitter posts and the flows estimated by a gravity model. More recently, McNeill, Bright, & Hale (2017) determined that estimated Twitter commuting flows outperform the radiation model, especially for short trips with higher volume of commuters. Kim, Park, & Lee (2018) compared different possible predictors in a gravity model of inner-city traffic. They used the resident population, the number of employees, and the number of tweets as proxies of mass values and found the number of tweets to be the most powerful predictor. Employing Foursquare user check-in data, Yang et al. (2015) combined clustering, regression, and gravity models to estimate an OD matrix of non-commuting trips in the Chicago urban area. The study concluded that the estimated OD matrix is similar to the ground truth OD flow matrix. Beiró et al. (2016) integrated Flickr data with a standard gravity model under a stacked regression procedure. The results showed that the hybrid gravity model outperforms the traditional gravity model. The research validated the performance of the model using two ground truth datasets of air travel and daily commuting in US counties. Utilizing Twitter data, Pouebrahim et al. (2018) compared gravity and neural network models to predict the commuter trip distribution in New York City. While these models had low predictive power, the findings indicated that adding Twitter data enhanced the performance of both models. The promising results achieved in these studies show the potential for enhancement in traditional modeling with social media data. However, more research is needed to fully grasp the scope of the predictive value of social media in estimating mobility patterns at fine spatial granularity. Applying new techniques and integrating traditional and new datasets are the first steps towards this goal.

4.2.3 Artificial Neural Networks and Decision Trees

Artificial neural networks (ANNs) started to be used in transportation research, including travel demand modeling, as an alternative to more conventional generalized linear models (GLM) and other econometric techniques from the beginning of the 1990s. Because ANNs have an inherent and demonstrated capability to capture nonlinearities and to be robust to alternative distributional properties of the data, they are found attractive for policy and planning analysis (Golshani et al., 2018). The usefulness of ANNs in trip distribution analysis and forecasting has been a focus of research, but reported results remain mixed. Black (1995) compared a gravity model and ANNs for a three-region flow problem and a nine-region commodity flow problem. Better performance of ANNs was reported for flow prediction in comparison to the gravity model. Similarly, Celik (2004) showed that ANNs outperform a regression model in predicting short-term inter-regional commodity flows. However, Tillema et al. (2006) reported that for a fifteen-region trip distribution in Rotterdam Rijnmond, ANNs perform better than gravity models only when data are limited. In another study, Mozolin et al. (2000) compared the performance of ANNs to the gravity models for commuter trip distribution among counties of the Atlanta Metropolitan Area, as well as among census tracts. The study suggested that although ANNs may fit the data better, their predictive accuracy is poor due to uncontrollable over-fitting and lack of generalization power, which cast doubt on the longitudinal transferability of ANN forecasts. Similar results were observed by Pouebrahim et al. (2018) who investigated commuting trip distribution between the census tracts of New York City.

Empirical results show the uncertainty about the potential contribution of ANNs in the context of trip distribution. In addition to problems of over-fitting and lack of generalization (Mozolin et al., 2000), the perception of ANNs as a "black box" makes it difficult to understand and utilize the results for planning purposes (Tillema et al., 2006). Decision trees are another class of techniques that have been used to model discrete decisions in travel demand. Thill & Wheeler (2000a , 2000b) demonstrated the merit of decision tree induction learning of spatial choice behavior in Minneapolis-St. Paul. Pitombo, de Souza, & Lindner (2017) compared these models to traditional gravity models of trip distribution in Bahia, Brazil, and concluded they exhibit better accuracy when prediction of destination choices is assessed by aggregate metrics such as trip length distribution and goodness-of-fit measures. More recent tree-based ensemble methods have been widely used in various machine learning problems due to their simplicities and understandability. Decision tree ensembles have shown promising results in studying travel time prediction (Zhang & Haghani, 2015), trip generation (Rashidi & Mohammadian, 2011), and mode choice modeling (Rasouli & Timmermans, 2014). As a specific class of decision tree ensembles, random forests (RFs) have been applied in predicting traffic flow (Leshem & Ritov, 2007) and travel time (Hamner, 2010), but only a few studies have applied the technique in travel demand modeling. Rasouli and Timmermans (2014) and Sekhar, Minal, & Madhu (2016) identified that RF outperforms other methods in mode choice modeling. Similarly, Ghasri et al. (2017) reported that RF shows high accuracy in estimating the total number of trips and trip attributes in a tour of trips at a disaggregate individual level. Following this research paradigm;

we explore the performance of RF compared to the gravity and ANN models in trip distribution modeling.

4.3 Methodology

We have selected New York City (NYC) as our study area (Figure 4.1) due to the large volume of readily available Twitter data. We focused on commuting trips because they are temporally stable and account for the largest share of total flows in a population (Yang et al., 2015). The census tracts are used as the geographic units for modeling commuting flows in NYC.

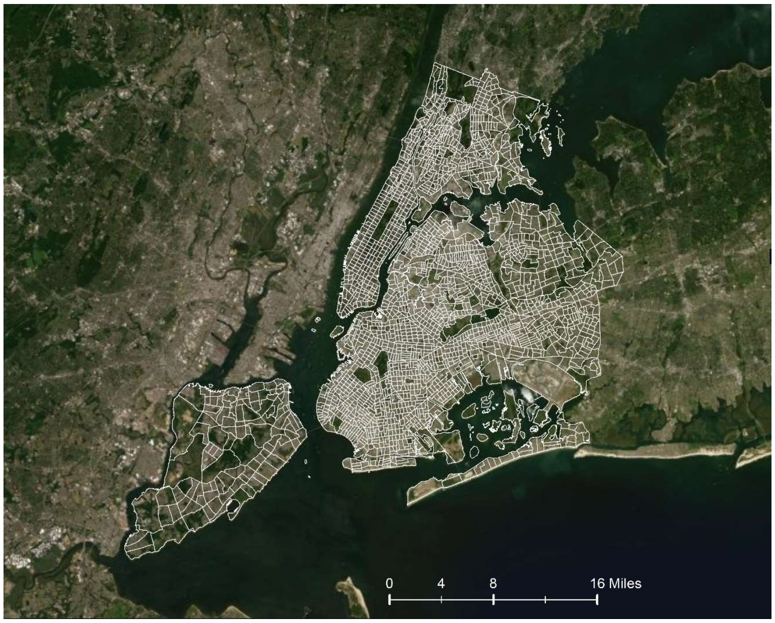


Figure 4. 1. Study Area (NYC Census Tracts).

4.3.1 Data Collection and Processing

The 2015 LEHD Origin-Destination Employment Statistics (LODES) for NYC were obtained from U.S. Census Bureau as the mobility variable of interest. The dataset reports the home and employment locations of workers, along with other characteristics such as age, earnings, industry distributions, and local workforce indicators. For each employee there is a home and a work census tract representing one commuting flow. The data were aggregated to census tract-to-tract commuting flows. The internal commuter flows, where home and work census tracts of commuters are the same, were excluded. In total, there were 903,685 origin-destination (OD) dyads and 2,580,596 home-work flows between census tracts in NYC that provide the dependent variable of this study. The number of OD dyads was further reduced to 878,132 after removing missing values on the input variables.

The key independent variables were selected based on previous commuting research (i.e., Sultana & Weber, 2014, 2007). These variables include residential population, employment, household median income, household median size, and household median number of vehicles. These data were obtained for the year 2015 from SimplyAnalytics, a socio-economic data provider. Two other input variables, points of interest (POI) and sprawl index for each census tract, were collected from the NYC OpenData and National Cancer Institute, respectively. The POIs are amenities (i.e., shops, tourist attractions, etc.) that different city agencies consider to be a common place or place/point of Interest. The sprawl index is a measure of compactness of a census tract based on various socio-economic, land use, and street network data (Sultana, Pourebrahim, & Kim, 2018; Ewing & Hamidi, 2014). We calculated the network distance between the centroids of origin and destination census tracts as another independent variable in the dataset. The North America Detailed Streets dataset was used to create the network data and all the procedures and calculations were performed in ArcGIS Network Analyst environment.

Twitter is among the most popular social media platforms that has been successfully used in the past studies (i.e., Pouebrahim et al., 2018). Approximately 700 million tweets are posted on Twitter per day by 126 million daily active users (Internet Live Stats, 2019), making it an optimal data source for collection of information including texts, pictures, and geolocation (Pourebrahim, Sultana, Edwards, Gochanour, & Mohanty, 2019). The large volume of readily available Twitter posts in NYC allowed us to conduct this research. Geolocated tweets posted within NYC from

June 2015 to May 2016 were obtained from the SOPHI data lake maintained by the Data Science Initiative (DSI) at the University of North Carolina at Charlotte. The dataset includes 1% sample of all tweets generated on Twitter during the timeframe. The tweets with precise location (longitude and latitude) were used, which resulted in 2,052,599 usable tweets. We then calculated the number of tweets and unique individual users in origin and destination census tracts. We kept the unique individual users (Twitter population) in the final analysis due to the high correlation (r=0.93, p < .05) between the two variables. In addition, multicollinearity was tested by computing the variation inflation factor (VIF) of the independent variables. A VIF of<2 for all the independent variables confirmed the absence of multicollinearity.

We extracted Twitter flows as another input variable by converting the Twitter dataset into an origin-destination matrix where origin represents home census tract and destination represents work census tract. Three filters were applied to the dataset: 1) Only users who posted geolocated tweets in two or more different census tracts were included; 2) The average length of stays for NYC international and domestic tourists is 9 and 1.9 days, respectively (Josephs, 2017). Therefore, Twitter users with a time interval of<10 days between their first and last tweets were considered tourists and removed from the dataset; and 3) Oxford Internet Institute identifies Twitter users with an average of>50 or 100–250 posts per day as bots or cyborgs, respectively (Nimmo, 2019). We used daily posting rate of 50 as the cutoff to remove suspicious bot activities. We also manually reviewed the "source" field ("generator" in our dataset) in the collected tweets following the work of Tsou et al. (2017) to identify the bots or cyborgs with commercial purposes. For example, if a tweet was created on an iPhone device, the source field will be "Twitter for iPhone". However, an advertisement tweet could have "TweetMyJOBS" or "dlvr.it" in its source field (Tsou et al., 2017). These noises were classified as advertisement, traffic, news, weather, and job. The total of 117,190 tweets were removed based on 15 identified source of noises that include dlvrt.it, pinprick on iOS, Squarespace, COS App, kickalert, Beer Menus, dine here, Dance Deets, 511NY-Tweets, TTN NYC traffic, Cities, eLobbyist, iembot, TweetMyJOBS, and SafeTweet by TweetMyJOBS.

The one-year timeframe of Twitter dataset allowed us to extract sufficient geolocated tweets of individuals in order to identify their home and work locations. The home and work census tracts for each user were identified and the direction of commuter flow was considered from home to work census tract. The home-work flows then were summed for all individuals having similar home-work locations. We identified the home and work census tracts of the Twitter users based on the common method of frequency counts with temporal (day-night) filtering (i.e., Jiang, Li, & Ye, 2018; McNeill et al., 2017). The most visited census tract at night (00:00–7:00) was considered the home location. If the frequency of visits to different census tracts were similar, the centroid of all the census tracts (represented as points) with the highest similar number of tweets was calculated to identify the home census tract. The most visited census tract during the day (8:00–17:00) excluding weekends and national holidays (Thanksgiving and Christmas) was considered the work location. Similarly, if the frequency of visits to different census tracts was the same, centroids were identified as the work census tract. The home census tracts for 9076 Twitter users and work census tracts for 7268 Twitter users were identified by centroid calculation. All other locations were assumed to be areas visited which are unrelated to either living or working. The users whose home and work locations could not be identified were removed from the analysis resulting in a total of 31,820 Twitter users. Internal flows in census tracts were also excluded from the dataset, which retained 31,688 home-work flows in the final dataset. These flows were then aggregated at census tract level to identify census tract-to-tract Twitter flows.

4.3.2 Trip Distribution Modeling

There was a high volume of zero Twitter flows in many census tracts in our dataset; we, therefore, first developed the three models only for those ODs (7445 ODs) that have both LODES flows and Twitter flows. These models were developed once without Twitter data, and then Twitter flow was added as a separate independent variable to the models. Since Twitter flow did not improve the models and the mean square errors (MSE) were large (see section 4), we dropped the variable in our final analyses. The final models were developed for 878,132 OD dyads first without the Twitter data discussed in the previous section and then with the addition of Twitter population in home and work census tracts. The next sections present the developed models, based on our final dataset that includes both non-Twitter data and Twitter population.

4.3.2.1 Gravity model

The gravity model used here can be formulated as follows (Eq. (1)):

$$T_{ij}=G\,\frac{M_i^{\alpha}M_j^{\beta}}{f(d_{ij})}$$

Where T_{ij} is the flow between two areas *i* (origin) and *j* (destination). In this study, the origin and destination are home and work census tracts and T_{ij} is the total number of home-work flows (commuting flows). M_i and M_j are trip production and attraction factors, respectively. $f(d_{ij})$ is the distance decay function commonly represented as a power function $f(d_{ij}) = M_{ij}^{-Y}$. In the original model, M_i and M_j are population of origin and destination. However, the trip production and attraction factors can be extended to other socioeconomic factors. Here, residential population, household median income, household median size, household median number of vehicles, and Twitter population at home census tract were used as trip production factors; also, employment, POI, sprawl, and Twitter population at work census tract served as trip attraction factors. The production and attraction factors and network distance between the centroids of census tracts were used to estimate the interzonal home-work flows in the gravity model. The model can be adjusted using a linear regression in the logarithmic scale (i.e., Beiró et al., 2016). Therefore, the final model is a linear combination of all variables in a log-log form. We quantified the estimation and compared the gravity model's performance with the ANN and RF models using the mean square error (MSE) and the coefficient of determination (R^2).

4.3.2.2 Artificial neural networks (ANNs)

ANNs are computational models that learn from and recognize patterns in data by mimicking the structure and functions of the nervous system in the brain (Nielsen, 1987). The fundamental building block of neural networks is the single-input neuron with three processes including the net input function, the weight function, and the transfer function. The output of a neuron with n inputs is calculated as follows (Eq. (2)) (Beale, Hagan, & Demuth, 2015):

$$a=f\sum_{i=1}^{n}(w_i p_i+b)$$

Where, a is output, p_i is input value, w_i is the weight, b is the bias and f is a transfer function of the neuron.

A widely used ANN topology is the multi-layer perceptron (MLP) (Figure 4.2) (Rumelhart & Mcclelland, 1986). It has been used in approximately 70% of ANN studies (de Oña & Garrido, 2014). The basic model of an MLP has three layers including input, hidden, and output. Each layer consists of neurons that are interconnected by weighted links that perform parallel distributed processing to solve a problem. The number of neurons in the input and output layers is usually determined by the number of predictor and predicted variables in the model (Amita, Singh, & Kumar, 2015). By adjusting the links' weights, a neural network learns the correlation between input and output. The learning process is much like a reward and punishment process so that when a desired/undesired output is generated, the weights related to the input are strengthened/reduced (Ding et al., 2013). Back-propagation (BP) is the most widely used algorithms in the learning process. We developed BP neural networks in MATLAB to predict the home-work flows using the same data used in the gravity model.

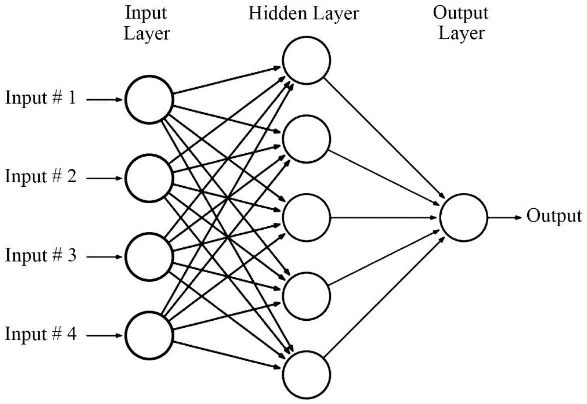


Figure 4. 2. Multi-Layer Perceptron (MLP) Structure.

The network has one hidden layer with 10 neurons in the input layer and one neuron in the output layer, corresponding to the number of input and output variables. The number of hidden layers and their neurons could be different depending on the complexity of connections between the input and output layers. However, a neural network with one hidden layer is a universal function approximator (de Oña & Garrido, 2014). Therefore, we used one hidden layer when developing the ANN model. It is common practice to select the number of hidden neurons by trial and error (Amita et al., 2015). We tested networks with 10 (number of inputs), 20, and 50 hidden neurons. Networks of larger size are impractical because of the excessive computational requirements for their training (Mozolin et al., 2000). Since no improvement in fit was observed with more hidden neurons in the model, we used 10 neurons in the final analysis. We employed two common transfer functions, the log-sigmoid and the linear for the hidden and output layers, respectively, and the Levenberg-Marquardt algorithm for training. The data were randomly divided into 70% for training, 15% for testing, and 15% for validation; the network was trained so that the MSE is minimized. Training was used for fitting and selecting the model, testing was used for evaluating the 'model's forecasting ability, and validation was used for determining the endpoint for the training process (minimizing the error) and avoiding overfitting (Srisaeng & Baxter, 2017).

4.3.2.3 Random forests (RFs)

The RF algorithm proposed by Breiman (2001) has been used for regression and classification, as well as for variable selection (Sulaiman, Shamsuddin, Abraham, & Sulaiman, 2011). An RF (Figure 4.3) is an ensemble of independent decision trees growing in parallel on a sub-sample of the training data (Lagomarsino, Tofani, Segoni, Catani, & Casagli, 2017). RF has a higher degree of accuracy compared to single trees, is effective in prediction, do not overfit, and allows measuring variable importance (Breiman, 2001; Lagomarsino et al., 2017; Sulaiman et al., 2011). The RF procedure includes (1) bootstrap resampling, (2) random variable selection, (3) out-of-bag error estimation, and (4) full depth decision tree growing (Ahmad, Mourshed, & Rezgui, 2017). RF is an ensemble of C trees $T_1(X)$, $T_2(X),\ldots, T_C(X)$, where $X = x_1, x_2,\ldots, x_m$ is an m-dimension vector of inputs. The output results for the trees (T_1 to T_C) are $\hat{Y}_1 = T_1(X)$, $\hat{Y}_2 = T_2(X),\ldots, \hat{Y}_C = T_C (X)$. The average value of all trees' outputs ($\hat{Y}_1$ to $\hat{Y}_C$) will be the final prediction $\hat{Y}$. An RF generates C number of decision trees from N training samples (Ahmad et al., 2017). For each tree in the forest, the algorithm takes a random sample of observations with replacement from the data and uses a random subset of the predictors at each split (Sulaiman et al., 2011). The model is built from 2/3 of the data (in-bag) and excluded data named out-of-bag (OOB) are used for identifying the prediction error (Ostmann & Martínez Arbizu, 2018b). The

OOB data eliminates the need for a separate test set for an unbiased estimate of error and can be used to identify the importance of predictors (Breiman, 2001). The relative importance of variables is identified by random permutation of out-of-bag data across each input variable to estimate the increase in the out-of-bag error (Breiman, 2001).

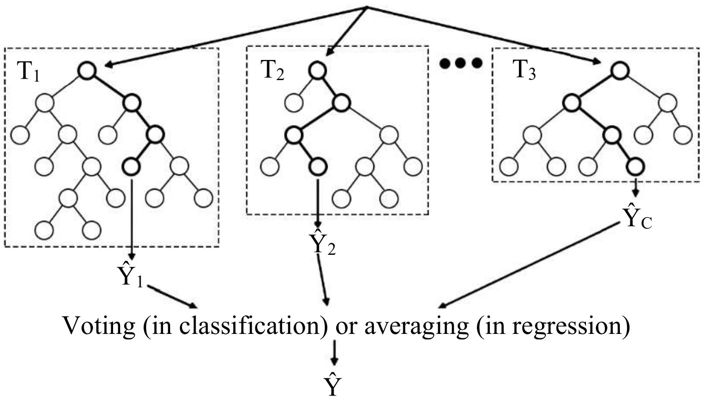


Figure 4. 3. Random Forest Structure.

We employed TreeBagger in MATLAB to generate the RF model. Three parameters need to be initialized in RF: (1) the minimum number of observations per tree leaf, (2) the number of randomly selected variables for each decision split, and (3) the number of trees. We trained a random forest with different leaf sizes of 5, 10, 20, 50, and 100. With the leaf size of 5 resulting in the minimum out-of-bag MSE, we set the minimum leaf size to 5. To develop an efficient forest, only a random subset of input variables is selected in RF to find the best splits (Ghasri et al., 2017). The Treebagger considers one third of the number of input variables for regression in each split. We used the same default value. Since increasing the number of trees in RF does not result in overfitting (Breiman, 2001), we trained random forests with 20, 50, 100, and 200 trees. No improvement was observed in the performance (decrease in OOB mean square error) of models with>100 trees. Therefore, we used 100 trees in the final model.

4.4 Results

Oure initial analysis was performed based on 7445 ODs that were represented by both LODES flows and Twitter flows. The gravity, ANN, and RF models were developed first with a specification that excludes Twitter data but includes variables: (1) network distance between ODs; (2) population, household median income, household median size, and household median number of vehicles in origin census tract; and (3) employment, sprawl, and POIs in destination census tract. Then we added the Twitter flow to the specification of each of the three models. The MSE and R2 for the six models are reported in Table 4.1.

Table 4. 1 Performance of the Gravity, ANN, and RF Models.

	Model	**MSE**	**R^2**
7445 ODs	Gravity without Twitter Data	110.23	0.25
	Gravity with Twitter flow	106.10	0.28
	Neural Network without Twitter Data	64.37	0.54
	Neural Network with Twitter flow	66.35	0.54
	Random Forest without Twitter Data	36.60	0.76
	Random Forest with Twitter Flow	37.78	0.75
878,132 ODs	Gravity without Twitter Data	17.33	0.09
	Gravity with Twitter Population	17.23	0.09
	Neural Network without Twitter Data	10.04	0.47
	Neural Network with Twitter Population	7.78	0.58
	Random Forest without Twitter Data	4.51	0.76
	Random Forest with Twitter Population	**4.26**	**0.78**

Note. The results for the gravity models are reported at non-logarithmic scales. Bold values indicate the best performance.

Contrary to past studies (i.e., Mozolin et al., 2000), the findings indicate the poor performance of the gravity model compared to the ANN and RF models in terms of both MSE and R2. The lowest MSE and highest R2 were achieved by the RF model, suggesting that a higher percentage of the home-work flows can be predicted by the input variables in the RF model with a lower error. Adding Twitter data to the ANN and RF not only did not improve the models, but also increased the MSEs for 1.98 and 1.18 respectively. However, the gravity model showed a 3% increase in the R2 and a decrease of 4.13 in MSE when adding the Twitter flows. These results are in contrast with current literature (i.e., Beiró et al., 2016; McNeill et al., 2017) that reported that flows extracted from social media such as Flickr or Twitter can be a good proxy to ground truth data or can outperform traditional models. Utilizing the relative importance of variables in the RF model, Twitter flow was among the least important variables (Figure 4.4). Past studies utilizing Twitter flows have been conducted on coarse spatial granularity, such as county. Our analysis is conducted on census tracts, which may explain that the volume of Twitter flows at this geographical scale is not significant enough to influence the performance of models.

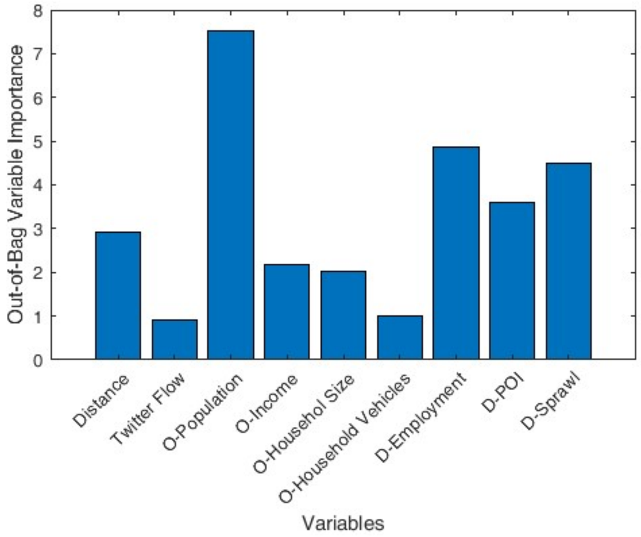


Figure 4. 4. Relative Importance of Variables in the RF Model for the 7445 OD Dyads.

Since Twitter flow did not improve the performance of the models in a meaningful way and the MSE for models were quite large, the variable was dropped from further analysis, which increased the OD dyads to total of 878,132. First, for comparison's sake, we developed the models for these OD dyads excluding Twitter data from the specification. The Twitter population in origins and destinations were then added to the model specifications. The performance of the gravity model remained similar to the previous results (the lowest R2 and highest MSE), while the RF model showed the best performance (Table 4.1). ANN and RF performances after adding Twitter population are illustrated in Figure 4.5. Predicted home-work flows have large errors in the initial phase of both models due to the small sample size, but errors decrease as the number of iterations increases. Although the performance of both test and validation sets in the ANN model are similar to the training set (no overfitting), the ANN training model does not fit the input data. The MSE for the ANN model is much higher than for the RF model, suggesting the shortcomings of ANNs for trip distribution modeling even with larger datasets.

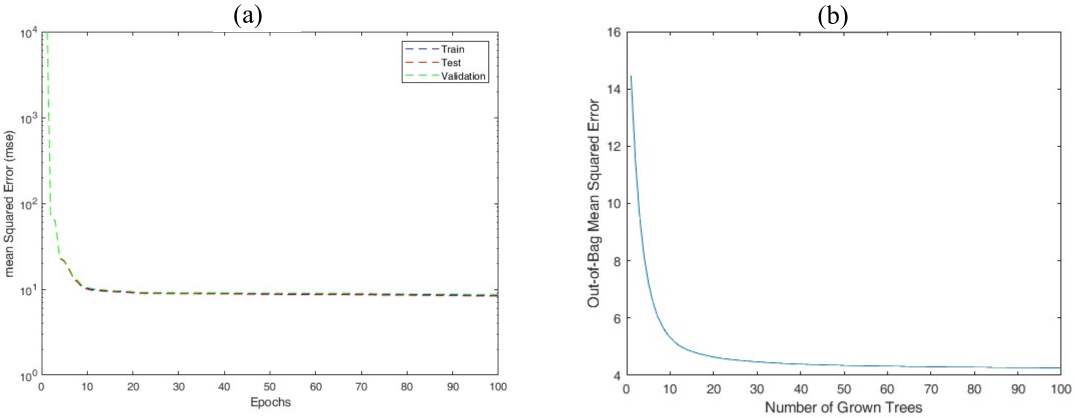


Figure 4. 5. Neural Network (a) and Random Forest (b) Performance (Including Twitter Population).

Although the ANN model did not have a good performance, the largest improvement was observed in the ANN model when Twitter population was added. The inclusion of Twitter population increased the predictive power of the ANN and RF models by 11 and 2%, respectively. The MSE decreased from 10.04 to 7.78 for the ANN model and from 4.50 to 4.26 for the RF model. The R2 of the gravity model did not change with the addition of Twitter population, but the MSE value slightly decreased by 0.10.

Figure 4.6 shows the difference between the actual and predicted numbers of incoming flows per census tract (summation of incoming flows to a census tract from all other tracts) for the gravity, ANN, and RF models developed with Twitter population. While both the gravity model and the ANN predictions show limited errors over large swaths of the city, the gravity model produces clusters of large underestimations (such as in Manhattan and Staten Island) but no extreme overestimation (Figure 4.6a), and the ANN model exhibits both extreme overestimations and underestimations in rather isolated pockets spread across the city, particularly in Manhattan (Figure 4.6b). The RF model has the most accurate estimation across the study area (Figure 4.6c).

One important note is that model performances are different depending on the size of the dataset. Comparing the models developed without Twitter data for 7445 and 878,132 OD dyads, all models showed an improvement in performance with a sharp decrease in MSE when the size of training data increased. The results for the ANN model are in contrast with past studies (i.e., Tillema et al., 2006), where ANNs were identified to be more reliable than the gravity models when data is scarce. These results suggest a general improvement in all models with additional data. The RF results are noteworthy from three perspectives: the models' predictive power, MSE, and the importance of predictor variables in explaining the output variable of home-work flows. Past studies of travel demand have reported traffic flows on the roads with an R2 ranging from 0.50 to 0.75 (Apronti & Ksaibati, 2018). Although our study is about trip distribution between census tracts, the resulting R2 of 0.77 can be regarded as a very good fit, especially at fine spatial granularity. The observed MSE is also lower compared to similar studies (i.e., Mozolin et al., 2000).

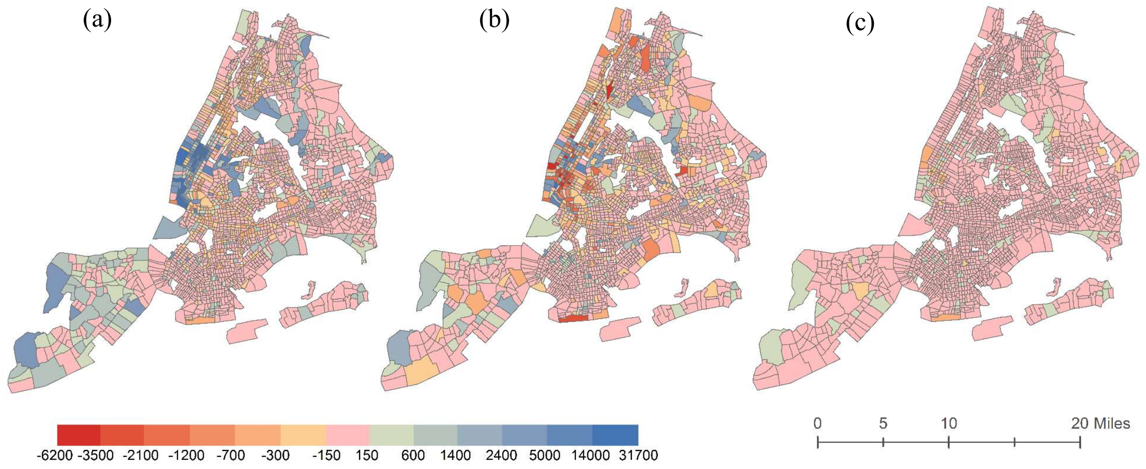


Figure 4. 6. Commuting Flows Modeling Error. These maps show the difference between the ground-truth and predicted numbers of incoming flows per census tract for the gravity (a), ANN (b), and RF (c) models. Blue marks underestimation by the models, red overestimation by the models, and pink the most accurate predictions.

To identify the relative importance of variables in the magnitude of OD flows, the model measures how much worse the MSE becomes after permutation of out-of-bag observations across each input variable. The larger the change, the more important the variable. Population of origin census tract was identified as the most important variable contributing to the model performance in both cases whether Twitter population is included or excluded (Figure 4.7). Destination employment, destination POI, destination sprawl index, and distance were among the other important variables when developing the RF model in the absence of Twitter data (Figure 4.7b). With the addition of Twitter population, the model identified the Twitter population of destination as one of the important variables (Figure 4.7a). We also developed an RF model using only the four most important variables (Figure 4.7c). The findings show that, with a limited number of input variables, the destination Twitter population becomes the most important factor contributing to the model.

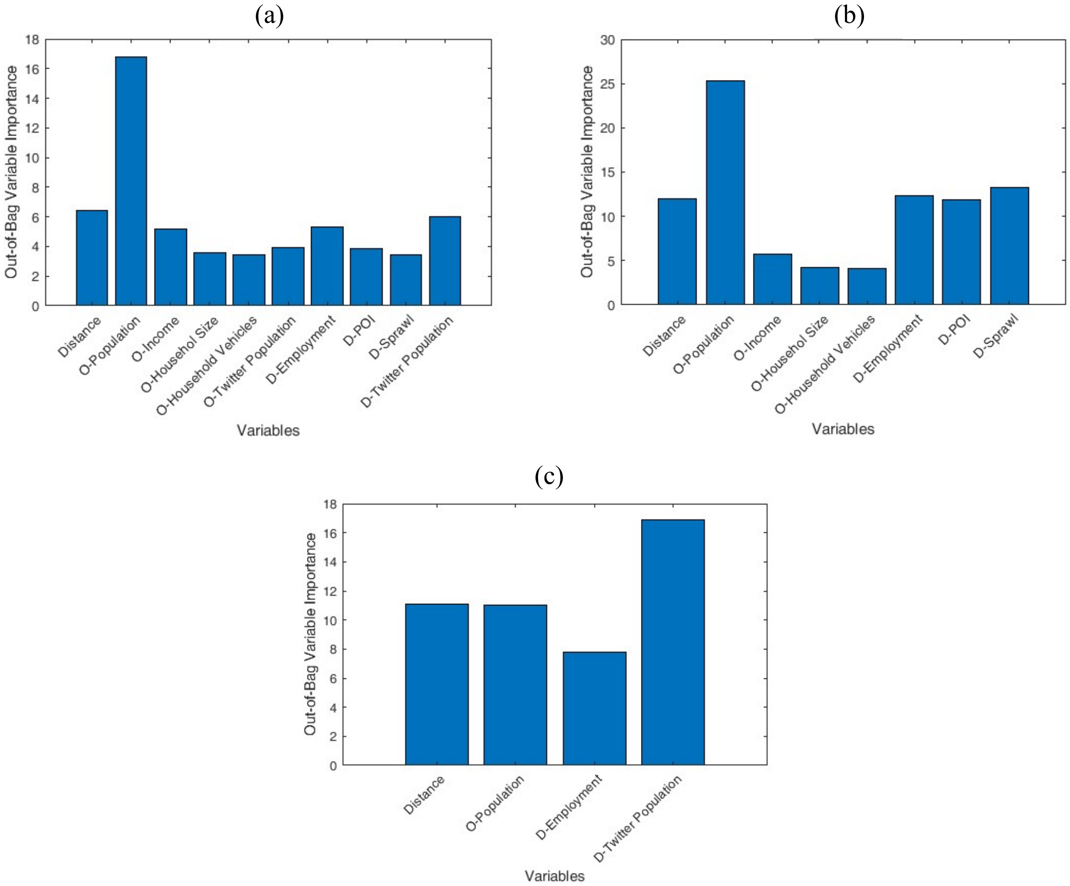


Figure 4. 7. Relative Importance of Variables for the RF Models Including Twitter Population (a), Excluding Twitter Population (b), and with the Four Important Variables (c).

4.5 Discussion

We compared the performance of the gravity, ANN, and RF models in commuting trip distribution at the census tract level in NYC. RF is identified as the best model, with the highest R2 and the lowest MSE. Travel demand analysis studies have commonly used statistical methods to model different travel components (i.e., travel mode, departure time, and trip destination) (Golshani et al., 2018). With the limitation of these methods in capturing the nonlinearities in the data, machine learning methods such as ANNs have gained popularity in transportation research. In trip distribution modeling, studies have indicated that ANN could provide similar goodness of fit as a gravity model, yet the error is usually higher for the ANN models (Yaldi, Taylor, & Yue, 2011, 2009; Tillema et al., 2006; Mozolin et al., 2000). Our results show better performance of ANN models with higher R2 and lower MSE compared to the gravity models. While Tillema et al. (2006) found better performance of ANNs (in terms of lower error) in case of data scarcity, our findings indicate a lower MSE for ANN models in both datasets (7445 and 878,132 ODs).

Despite the lower performance of ANNs in trip distribution modeling, there have been other studies pointing to promising results of ANNs for modeling traffic flows. In short-term traffic forecasting, ANNs (individual or hybrid) have mostly outperformed other methods (i.e., (Stathopoulos, Dimitriou, & Tsekeris, 2008; Vlahogianni, Karlaftis, & Golias, 2007). The primary focus in these studies was the time-series prediction to test the accuracy against the traditional models. One can conclude that ANNs might show better performance when dealing with temporal aspects of the traffic data. Therefore, ANNs could be more appropriate in developing dynamic models. Developing time-series trip distribution models might be a good practice for future research.

Trip distribution modeling is a complicated and important step in transport planning. The errors generated during this step will pass on to the other steps (Tillema et al., 2006). Developing more accurate models, therefore, is critical for planning purposes such as alleviating traffic congestion problems. While RF model had the best performance in this study, the ANN model also showed a lower MSE compared to similar studies (i.e., Mozolin et al., 2000). Past studies have mostly been conducted at coarse spatial granularity such as county that dealt with much fewer zones compared to our study. A study similar to ours was conducted by Mozolin et al. (2000) who modeled trip distribution at the census tract level. The MSE observed for the ANN model in our research was lower compared to this earlier study. The better performance of our ANN model might confirm that the appropriate selection of the variables is important in trip distribution modeling. Despite the better performance of ANN model compared to previous studies, its predictive accuracy is poor on the whole scale. The RF model not only resulted in a higher goodness of fit, but also a lower MSE. While RFs have been identified among the best ensemble methods for modeling trip generation and modal split (Ghasri et al., 2017; Rasouli & Timmermans, 2014), our results confirm their suitability in trip distribution modeling as well. In addition, RF allowed us to identify the importance of variables in the model.

Contrary to past research, Twitter flow was not an important variable in the RF model. It is our conjecture this may be due to the fine spatial granularity at which we modeled trip distribution. Past studies (McNeill et al., 2017; Beiró et al., 2016) at the county or larger scales have identified Twitter flows as a good proxy for mobility flows. A more detailed analysis for Twitter flows should be conducted probably in specific areas, where there are more flows. The RF model showed a slight improvement with the addition of Twitter population. Twitter population at destination (work census tract) was the third most influential variable after origin (home census tract) population and distance. Its importance even increased when developing the model with fewer variables. Hence, Twitter population can be a good predictor of trip distribution when data on more conventional predictors are scarce, incomplete, outdated or uncertain. This would be useful in cases where the major aim is a higher accuracy and better prediction for modeling trip distribution. When Twitter population was dropped from the model specification, sprawl index and POI were identified among the significant variables. This is particularly meaningful when the objective is to understand the mechanism behind the modeling process and identifying major factors. From a policy analysis perspective, it is essential to identify how specific factors such as land use can affect trip distribution within cities.

Twitter population may in fact be a proxy measure of attraction of a census tract, where there are higher number of points of interest and therefore, higher activities. Higher activities are usually happening at denser areas, where mixed land uses attract more population. Therefore, Twitter as a proxy for the attractiveness of these areas can be translated to a trip attraction factor in modeling trip distribution. Twitter has the potential to be a good source of data to develop population dynamics that can be used for planning purposes. These data could be instrumental in better apprehending critical attraction locations in cities and arranging sustainable transportation alternatives (Yang et al., 2015). Additionally, Population dynamics can be applied to evacuation planning models to forecast movements in real time. With the increasingly adverse effects of climate change such as sever hurricanes, such data are becoming more valuable for better urban management (Eshghi & Schmidtke, 2018). There would also be more potential to conduct dynamic travel demand modeling and monitoring in the near future with the increase in the number of people using social media. The use of these data coupled with artificial intelligence may lead to smart and sustainable solutions for travel demand planning and management.

4.6 Conclusions

A primary aim of transportation policy makers is to achieve sustainable mobility in urban areas (Kepaptsoglou et al., 2012; May, 2013). However, collecting travel demand data at high spatio-temporal resolution is the major gap that exists between the current state of the practice and an efficient urban sustainable solution (Yang et al., 2015). Social media may be a useful source of data to achieve this goal, yet their potential remains insufficiently investigated. The primary focus in our study was to explore the integration of social media data with traditional data in estimating trip distribution within cities and in comparing different models to identify the model with the highest performance. Given the identified need in literature for utilizing more factors in gravity modeling combined with new techniques, three models including gravity, neural network, and random forest were developed for predicting the home-work flows between census tracts of New York City. The results showed that the models performed better when the Twitter population enters the model specification as input. The random forests model showed the best performance. Twitter flow was not identified as an important factor in the models that would warrant further investigation. However, Twitter population, particularly at the work census tract, was a significant factor.

The findings suggest the potential for using social media for developing dynamic models of population and trip distribution in future research to achieve more sustainable solutions for cities. However, this research has several limitations. Twitter population is not totally representative of all demographic and socioeconomic groups within a population. In addition, not all Twitter users share their location, thus resulting in a population bias (Jiang et al., 2018) that may result in overestimations or underestimations in the model results. Therefore, these biases need to be considered when using volunteered geographic information (Jiang & Thill, 2015). Despite all these limitations, social media data are still valuable especially in areas with high population density (Li, Goodchild, & Xu, 2013). Developing more advanced methods to draw a representative sample from social media (Jiang, Li, & Cutter, 2019) remains an important issue for future research.

CHAPTER V
CONCLUSIONS

Big data provide new opportunities for scholars in various disciplines to understand human dynamics in a way that was not possible in the past. This dissertation, by integrating both theory- and data-driven approaches, studied human communications and activities using social media information system. An introduction to the research problem was provided in Chapter I, where I reviewed existing theories and concepts to setup the conceptual framework of this dissertation. Chapter II assessed communication dynamics on Twitter during Hurricane Sandy to address the conceptual link 1—human dynamics (communications), space (cyberspace), social media information system. Twitter usage during the hurricane was identified following the survey of the general population in the affected counties. Twitter users' discussions, sentiments, and networks were explored through generated tweets in the entire Twitter platform in addition to geotagged tweets in the affected areas. Geotagged tweet activities in NYC during Hurricane Sandy were used in Chapter III to identify events from both data and theory perspectives. This chapter first addressed the conceptual link 3—human dynamics (communication and activities), place, social media information system, which focused on generation of place by (1) integrating human communications and activities and (2) human movements. Human movements during Hurricane Sandy were then explored in the same chapter to address the conceptual link 4—human dynamics (activities), place, social media information system, which focused on generation of human activities by place with a modeling approach. Link 4 was also addressed in Chapter IV to examine the potential of Twitter in modeling commuting trip distribution in New York City census tracts.

A shift from analysis to syntheis has been identified crucial for an interdisiplenary field such as geography (Gober, 2000). While geography builds on spatial analysis, human-environment interaction, and place-based and regional analyses (Baerwald, 2010), the decipline lacks in theories. From a synthesis perspective, employing theories from other disciplines and contextualizing them with geographic concepta is an important step. With the current shift of geographic research from a data-scarce to a data-rich environment (Miller & Goodchild, 2015), the primary criticism of such data is their non-theoretical approach in deducing meaning (Shelton et al., 2014). Developing theoretical frameworks, therefore, is essential to achive the full potential of such data in geographic context. To address this challenge, this dissertation developed a platform for research on effective use of social media utilizing existing information systems (IS) theories and geographic concepts of space and place to understand human dynamics.

From the effective use theory perspective, the goal of social media system is to achieve an effective communication and human dynamics awareness. Regarding effective communication, the first study revealed the most important information that can be derived from twitter during disasters so that authorities can successfully utilize such data. However, a lack of bi-directional communications during Hurricane Sandy demonstrates that social media was not effectively utilized during the event. The second and third case studies reveal the potential of social media for human dynamic awareness with a data synthesis approach (Yang, 2015). The second study combined both geotagged information and qualitative text generated through Twitter to provide insights for enhancing situational awareness during natural disasters. Similarly, the third study identified potential of social media in transportation modeling by integrating both traditional and social media data and utilizing machine learning techniques.

Past studies have identified difficulties in deriving scientific results through social media due to data issues such as quality, credibility, and biased demographics of users (Zou et al., 2018). While social media data have mostly assessed through Rumor theory, the effective use theory presented in this dissertation offers new insights to the data issue. In their study of contextualized effective use, Burton-Jones & Volkoff (2017), identified accuracy, consistency, and reflection-in-action as three dimensions of affordance actualization. Although their study was conducted at a health organization system, same considerations can be applied in social media system. Accuracy in their study, for example, was identified as truth, nothing but truth, and whole truth. Truth and nothing but truth in social media analysis can be achieved through data cleaning and noise removal. Developing automated data-learned models to filter relevant information is a possible future direction. These models can benefit from inclusion of all possible information from social media data (i.e., text, user information, image) and geospatial data (i.e., precipitation, wind) to extract reliable data. Whole truth as identified by the effective use theory is another dimension that need careful consideration when applying such data. Social media data such as Twitter is not a full representation of the world. This issue was observed in the second and third case studies (Chapters III and IV) in this dissertation. For instance, despite many events and damages in different parts of NYC, the detected events were mostly concentrated in Manhattan, where tweet activities are generally higher. Similarly, Twitter flow was not representative of trip distribution at fine spatial granularity. One way to tackle these issues is data integration, where data from multiple sources such as different social media platforms and sensors in cities are used. However, this will pose additional methodological and computational challenges (Zou et al., 2018) that should be properly addressed in future research. Another limitation of social media data is their demographic and socioeconomic biases. In addition, not every user share his/her location, thus resulting in a population bias (Jiang et al., 2018). Therefore, these biases need to be considered when using such data.

The case studies here also bring our attention to the issue of scale and how it affects the “whole truth” dimension of accuracy in effective use theory. The scale at which social media should be analyzed largely depends on the topic being investigated. This dissertation highlights the need for analysis at coarse spatial granularity in the mobility research, while fine spatial scale analysis (i.e., census block) is more effective in identifying activities and events.

In the general model of effective use, Burton-Jones & Grange (2013), identified transparent interaction, representational fidelity, and informed action as the three dimensions of effective use. As discussed in Chapter I, transparent interaction refers to the extent to which users can access the system’s representation by its surface and physical structure; representation fidelity is the extent to which users obtain faithful representation of the domain; and informed action is the extent to which users act upon these representations to improve their state. By employing these dimensions in social media information system, specific tools that facilitate transparent interaction and help users in analyzing the data in a consistent manner is another possible future direction. These tools should be contextualized according the need of specific stakeholders such as disaster managers or urban and transportation planners. Examples of such tools are SMART dashboard and mobile app developed to monitor social behavior changes in different events using Twitter messages (Yang et al., 2016). However, a challenge with these tools is the need for frequent update in the backend development languages and the promotional strategy for using the tools and apps. Another challenge is the accuracy of the shared data. To achieve higher accuracy, these tools can facilitate the interaction of the user who is generating the data (i.e., Twitter user). For example, developing applications that can be linked to the common social media platforms and ask the social media user to share specific information (i.e., image) when posting about a particular event might help in identifying the accuracy of the shared information. Integration and assessment of these tools at the organizational level would be the next step to improve urban decision making. However, as a prerequisite of such assessments, a context specific theory of effective use of information systems needs to be first developed and tested from an urban perspective. True representation of the world will be the essential part of such systems to implement informed actions in cities.

REFERENCES

Abbas, A. k., Bayat, O., & Ucan, O. N. (2018). Estimation of Twitter user's nationality based on friends and followers information. *Computers & Electrical Engineering*, *66*, 517–530. https://doi.org/10.1016/j.compeleceng.2017.06.033

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(4), 433–459. https://doi.org/10.1002/wics.101

Acar, A., & Muraki, Y. (2011). Twitter for crisis communication: lessons learned from Japan's tsunami disaster. *International Journal of Web Based Communities*, *7*(3), 392. https://doi.org/10.1504/IJWBC.2011.041206

Agarwal, R. (2012). Editorial Notes. *Information Systems Research*, *23*(4), 1087–1092. https://doi.org/10.1287/isre.1120.0458

Agnew, J. A. (1987). *Place and Politics: The Geographical Mediation of State and Society*. Retrieved from https://www.taylorfrancis.com/books/9781315756585

Ahmad, M. W., Mourshed, M., & Rezgui, Y. (2017). Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption. *Energy and Buildings*, *147*, 77–89. https://doi.org/10.1016/j.enbuild.2017.04.038

Allen, C., Tsou, M. H., Aslam, A., Nagel, A., & Gawron, J. M. (2016). Applying GIS and machine learning methods to twitter data for multiscale surveillance of influenza. *PLoS ONE*, *11*(7), 1–10. https://doi.org/10.1371/journal.pone.0157734

Amita, J., Singh, J. S., & Kumar, G. P. (2015). Prediction of Bus Travel Time Using Artificlal Neural Network. *International Journal for Traffic and Transport Engineering*, *5*(4), 410–424.

Anda, C., Erath, A., & Fourie, P. J. (2017). Transport modelling in the age of big data. *International Journal of Urban Sciences*, *21*(sup1), 19–42. https://doi.org/10.1080/12265934.2017.1281150

Anderson, C. (2008). The end of theory: the data deluge makes the scientific method obsolete. *Wired Magazine*. Retrieved from http://www.uvm.edu/pdodds/files/papers/others/2008/anderson2008a.pdf

Ann St Denis, L., Palen, L., & Anderson, K. M. (2014). Mastering Social Media: An Analysis of Jefferson County's Communications during the 2013 Colorado Floods. *Proceedings of the 11th International ISCRAM Conference*, 737–746. University Park, Pennsylvania.

Anson, S., Watson, H., Wadhwa, K., & Metz, K. (2017). Analysing social media data for disaster preparedness: Understanding the opportunities and barriers faced by humanitarian actors. *International Journal of Disaster Risk Reduction*, *21*(November 2016), 131–139. https://doi.org/10.1016/j.ijdrr.2016.11.014

Apronti, D. T., & Ksaibati, K. (2018). Four-step travel demand model implementation for estimating traffic volumes on rural low-volume roads in Wyoming. *Transportation Planning and Technology*, *41*(5), 557–571. https://doi.org/10.1080/03081060.2018.1469288

Baerwald, T. J. (2010). Prospects for geography as an interdisciplinary discipline. *Annals of the Association of American Geographers*, *100*(3), 493–501. https://doi.org/10.1080/00045608.2010.485443

Barabási, A.-L. (2005). The origin of bursts and heavy tails in human dynamics. *Nature*, *435*(7039), 207–211. https://doi.org/10.1038/nature03459

Barbosa, H., Barthelemy, M., Ghoshal, G., James, C. R., Lenormand, M., Louail, T., … Tomasini, M. (2018). Human mobility: Models and applications. *Physics Reports*, *734*, 1–74. https://doi.org/10.1016/j.physrep.2018.01.001

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An Open Source Software for Exploring and Manipulating Networks. *Proceedings of the Third International ICWSM Conference*, 361–362. Retrieved from http://qualitysafety.bmj.com/lookup/doi/10.1136/qshc.2004.010033

Beale, M. H., Hagan, M. T., & Demuth, H. B. (2015). Neural Network Toolbox User ' s Guide. In *MathWorks*. https://doi.org/10.1002/0471221546

Beiró, M. G., Panisson, A., Tizzoni, M., & Cattuto, C. (2016). Predicting human mobility through the assimilation of social media traces into mobility models. *EPJ Data Science*, *5*(1), 30. https://doi.org/10.1140/epjds/s13688-016-0092-2

Berger, A. D. (2012). *A Travel Demand Model for Rural Areas*. (August).

Birant, D., & Kut, A. (2007). ST-DBSCAN: An algorithm for clustering spatial-temporal data. *Data and Knowledge Engineering*, *60*(1), 208–221. https://doi.org/10.1016/j.datak.2006.01.013

Black, W. R. (1995). Spatial interaction modeling using artificial neural networks. *Journal of Transport Geography*, *3*(3), 159–166. https://doi.org/10.1016/0966-6923(95)00013-S

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation David. *Journal of Machine Learning Research*, *3*, 993–1022. https://doi.org/10.1162/jmlr.2003.3.4-5.993

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

Bonaretti, D., & Piccoli, G. (2019). Unifying the Emergency Management Research Program in Is : A Representation Theory Perspective for Effective Use in Chaotic Environments. *Twenty-Seventh European Conference on Information Systems (ECIS2019)*, (April). Stockholm-Uppsala, Sweden.

Brady, A., Brookes, R., Brown, N., Brown, W., Perry, E., & Wilhite, C. (2013). *Situational Awareness*. Retrieved from http://www1.mwcog.org/ire/projects/IREproj4.pdf

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Burke, J. A., Spence, P. R., & Lachlan, K. A. (2010). Crisis preparation, media use, and information seeking during Hurricane Ike: Lessons learned for emergency communication. *Journal of Emergency Management*, *8*(5), 27–37. https://doi.org/10.5055/jem.2010.0030

Burton-Jones, A., & Grange, C. (2013). From Use to Effective Use: A Representation Theory Perspective. *Information Systems Research*, *24*(3), 632–658. https://doi.org/10.1287/isre.1120.0444

Burton-Jones, A., & Volkoff, O. (2017). How can we develop contextualized theories of effective use? A demonstration in the context of community-care electronic health records. *Information Systems Research*, *28*(3), 468–489. https://doi.org/10.1287/isre.2017.0702

Candia, J., González, M. C., Wang, P., Schoenharl, T., Madey, G., & Barabási, A. L. (2008). Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical*, *41*(22), 1–16. https://doi.org/10.1088/1751-8113/41/22/224015

Caragea, C. ., Squicciarini, A. ., Stehle, S. ., Neppalli, K. ., & Tapia, A. . (2014). Mapping moods: Geo-mapped sentiment analysis during hurricane sandy. *ISCRAM 2014 Conference Proceedings - 11th International Conference on Information Systems for Crisis Response and Management*, (May), 642–651. Retrieved from http://www.iscram.org/legacy/ISCRAM2014/papers/p29.pdf

Celik, H. M. (2004). Modeling freight distribution using artificial neural networks. *Journal of Transport Geography*, *12*(2), 141–148. https://doi.org/10.1016/j.jtrangeo.2003.12.003

Chatfield, A. T., & Brajawidagda, U. (2012). Twitter tsunami early warning network: a social network analysis of Twitter information flows. *23rd Australasian Conference on Information Systems*, 1–10.

Chatfield, A. T., & Reddick, C. G. (2015). Understanding Risk Communication Gaps through E-Government Website and Twitter Hashtag Content Analyses: The Case of Indonesia's Mt. Sinabung Eruption. *Journal of Homeland Security and Emergency Management*, *12*(2). https://doi.org/10.1515/jhsem-2014-0086

Chatfield, A. T., Scholl, H. J., & Brajawidagda, U. (2014). #Sandy Tweets: Citizens' Co-Production of Time-Critical Information during an Unfolding Catastrophe. *2014 47th Hawaii International Conference on System Sciences*, 1947–1957. https://doi.org/10.1109/HICSS.2014.247

Cresswell, T. (2014). *Place: an introduction*. John Wiley & Sons.

Cresswell, T., & Merriman, P. (Eds.). (2011). *Geographies of mobilities: Practices, spaces, subjects*. Ashgate Publishing, Ltd.

Croitoru, A., Wayant, N., Crooks, A., Radzikowski, J., & Stefanidis, A. (2015). Linking cyber and physical spaces through community detection and clustering in social media feeds. *Computers, Environment and Urban Systems*, *53*, 47–64. https://doi.org/10.1016/j.compenvurbsys.2014.11.002

Crooks, A., Croitoru, A., Stefanidis, A., & Radzikowski, J. (2013). #Earthquake: Twitter as a Distributed Sensor System. *Transactions in GIS*, *17*(1), 124–147. https://doi.org/10.1111/j.1467-9671.2012.01359.x

de Dios Ortuzar, J., & Willumsen, L. G. (2011). *Modelling transport*. John Wiley & Sons.

de Oña, J., & Garrido, C. (2014). Extracting the contribution of independent variables in neural network models: A new approach to handle instability. *Neural Computing and Applications*, *25*(3–4), 859–869. https://doi.org/10.1007/s00521-014-1573-5

DeLanda, M. (2006). *A New Philosophy of Society : Assemblage Theory and Social Complexity*. Retrieved from http://www.bloomsburycollections.com/book/a-new-philosophy-of-society-assemblage-theory-and-social-complexity

Ding, C., Wang, W., Wang, X., & Baumann, M. (2013). A Neural Network Model for Driver's Lane-Changing Trajectory Prediction in Urban Traffic Flow. *Mathematical Problems in Engineering*, *2013*, 1–8. https://doi.org/10.1155/2013/967358

Edwards, C., Spence, P. R., Gentile, C. J., Edwards, A., & Edwards, A. (2013). How much Klout do you have…A test of system generated cues on source credibility. *Computers in Human Behavior*, *29*(5), A12–A16. https://doi.org/10.1016/j.chb.2012.12.034

Endsley, M. R. (1988). Design and evaluation for situation awareness enhancemen. 97–101. *Proceedings of the Human Factors Society-32nd Annual Meeting.*

Eshghi, M., & Alesheikh, A. A. (2016). Introducing a Conceptual Model to Improve the Quality of Storage of Volunteered Geographic Information: In the Field of "Fitness-for-Use" Indicator. *Journal of Geomatics Science and Technology*, *6*(1), 19–32. Retrieved from http://jgst.issge.ir/article-1-393-en.html

Eshghi, M., & Schmidtke, H. R. (2018). An approach for safer navigation under severe hurricane damage. *Journal of Reliable Intelligent Environments*, *4*(3), 161–185. https://doi.org/10.1007/s40860-018-0066-1

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *2nd Inter National Conference on Knowledge Discovery and Data Mining*, *96*, 226–231. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/B9780444527011000673

Ewing, R., & Hamidi, S. (2014). *Measuring urban sprawl and validating sprawl measures*. Retrieved from file:///C:/Users/aamiraza/Downloads/sprawl-report-short (1).pdf

Feldman, D., Contreras, S., Karlin, B., Basolo, V., Matthew, R., Sanders, B., … Luke, A. (2016). Communicating flood risk: Looking back and forward at traditional and social media outlets. *International Journal of Disaster Risk Reduction*, *15*(December), 43–51. https://doi.org/10.1016/j.ijdrr.2015.12.004

Flanagan, B. E., Gregory, E. W., Hallisey, E. J., Heitgerd, J. L., & Lewis, B. (2011). A Social Vulnerability Index for Disaster Management. *Journal of Homeland Security and Emergency Management*, *8*(1). https://doi.org/10.2202/1547-7355.1792

Floridi, L. (2012). Big data and their epistemological challenge. *Philosophy and Technology*, *25*(4), 435–437. https://doi.org/10.1007/s13347-012-0093-4

Georganos, S., Grippa, T., Niang Gadiaga, A., Linard, C., Lennert, M., Vanhuysse, S., … Kalogirou, S. (2019). Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. *Geocarto International*, *0*(0), 1–16. https://doi.org/10.1080/10106049.2019.1595177

Ghasri, M., Hossein Rashidi, T., & Waller, S. T. (2017). Developing a disaggregate travel demand system of models using data mining techniques. *Transportation Research Part A: Policy and Practice*, *105*(June 2016), 138–153. https://doi.org/10.1016/j.tra.2017.08.020

Gnip APIs. (n.d.). Retrieved April 1, 2015, from http://support.gnip.com/apis/

Gober, P. (2000). In Search of Synthesis. *Annals of the Association of American Geographers*, *90*(1), 1–11. https://doi.org/10.1111/0004-5608.00181

Golshani, N., Shabanpour, R., Mahmoudifard, S. M., Derrible, S., & Mohammadian, A. (2018). Modeling travel mode and timing decisions: Comparison of artificial neural networks and copula-based joint model. *Travel Behaviour and Society*, *10*, 21–32. https://doi.org/10.1016/j.tbs.2017.09.003

González, M. C., Hidalgo, C. A., & Barabási, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, *453*(7196), 779–782. https://doi.org/10.1038/nature06958

Graham, M. W., Avery, E. J., & Park, S. (2015). The role of social media in local government crisis communications. *Public Relations Review*, *41*(3), 386–394. https://doi.org/10.1016/j.pubrev.2015.02.001

Grube, L., & Storr, V. H. (2014). The capacity for self-governance and post-disaster resiliency. *The Review of Austrian Economics*, *27*(3), 301–324. https://doi.org/10.1007/s11138-013-0210-3

Guskin, E., & Hitlin, P. (2012). Hurricane Sandy and Twitter. *Pew Research Center*. Retrieved from http://www.journalism.org/2012/11/06/hurricane-sandy-and-twitter/

Hamner, B. (2010). Predicting travel times with context-dependent random forests by modeling local and aggregate traffic flow. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 1357–1359. https://doi.org/10.1109/ICDMW.2010.128

Hanneman, R., & Riddle, M. (2005). *Introduction to Social Network Methods*. Retrieved from http://faculty.ucr.edu/~hanneman/

Hansen, D. L., Shneiderman, B., & Smith, M. A. (2010). Social Network Analysis: Measuring, Mapping, and Modeling Collections of Connections. In *Analyzing Social Media Networks with NodeXL : Insights from a Connected World* (pp. 31–49). Elsevier Science & Technology.

Hasan, S., Zhan, X., & Ukkusuri, S. V. (2013). Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing - UrbComp '13*, 1. https://doi.org/10.1145/2505821.2505823

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., & Ratti, C. (2014). Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, *41*(3), 260–271. https://doi.org/10.1080/15230406.2014.890072

Heymann, S., & Grand, B. Le. (2013). Visual Analysis of Complex Networks for Business Intelligence with Gephi. *2013 17th International Conference on Information Visualisation*, 307–312. https://doi.org/10.1109/IV.2013.39

Houston, J. B., Hawthorne, J., Perreault, M. F., Park, E. H., Goldstein Hode, M., Halliwell, M. R., … Griffith, S. A. (2015). Social media and disasters: A functional framework for social media use in disaster planning, response, and research. *Disasters*, *39*(1), 1–22. https://doi.org/10.1111/disa.12092

Huang, X., Li, Z., Wang, C., & Ning, H. (2019). Identifying disaster related social media for rapid response: a visual-textual fused CNN architecture. *International Journal of Digital Earth*, *0*(0), 1–23. https://doi.org/10.1080/17538947.2019.1633425

Huang, Y., Li, Y., & Shan, J. (2018). Spatial-Temporal Event Detection from Geo-Tagged Tweets. *ISPRS International Journal of Geo-Information*, *7*(4), 150. https://doi.org/10.3390/ijgi7040150

Hughes, A. L., St. Denis, L. A. A., Palen, L., & Anderson, K. M. (2014). Online public communications by police &amp; fire services during the 2012 Hurricane Sandy. *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems - CHI '14*, 1505–1514. https://doi.org/10.1145/2556288.2557227

Internet Live Stats. (2019). Retrieved April 29, 2019, from www. internetlivestats.com

Jackson, M. O. (2008). *Social and Economic Networks*. https://doi.org/10.1017/CBO9781107415324.004

Jiang, B., & Thill, J.-C. (2015). Volunteered Geographic Information: Towards the establishment of a new paradigm. *Computers, Environment and Urban Systems*, *53*, 1–3. https://doi.org/10.1016/j.compenvurbsys.2015.09.011

Jiang, Y., Li, Z., & Cutter, S. L. (2019). Social network , activity space , sentiment and evacuation : what can social media tell us ? *Annals of the Association of American Geographers*, (February).

Jiang, Y., Li, Z., & Ye, X. (2018). Understanding demographic and socioeconomic biases of geotagged Twitter users at the county level. *Cartography and Geographic Information Science*, *46*(3), 1–15. https://doi.org/10.1080/15230406.2018.1434834

Jones, J. P., Marston, S. A., & Woodward, K. (2011). Scales and Networks – Part II. *The Wiley-Blackwell Companion to Human Geography*, 404–414. https://doi.org/10.1002/9781444395839.ch28

Jones, K. T. (1998). Scale as epistemology. *Political Geography*, *17*(1), 25–28. https://doi.org/10.1016/S0962-6298(97)00049-8

Josephs, L. (2017, April). *New York City needs foreign visitors because they spend four times more money than Americans*. Retrieved from https://qz.com/954413/new-york-city-needs-foreign-tourists-because-they-spend-more/

Joshi, A., & Aoki, M. (2014). The role of social capital and public policy in disaster recovery: A case study of Tamil Nadu State, India. *International Journal of Disaster Risk Reduction*, *7*, 100–108. https://doi.org/10.1016/j.ijdrr.2013.09.004

Jung, J.-Y., & Moro, M. (2014). Multi-level functionality of social media in the aftermath of the Great East Japan Earthquake. *Disasters*, *38*(s2), s123–s143. https://doi.org/10.1111/disa.12071

Kaiser, R., & Nikiforova, E. (2008). The performativity of scale: the social construction of scale effects in Narva, Estonia. *Environment and Planning D: Society and Space*, *26*(3), 537–562. https://doi.org/10.1068/d3307

Kalogirou, S., & Georganos, S. (2019). Package 'SpatialML.' In *Geocarto International*. https://doi.org/10.1080/10106049.2019.1595177

Karduni, A., Cho, I., Wessel, G., Ribarsky, W., Sauda, E., & Dou, W. (2017). Urban space explorer: A visual analytics system for urban planning. *IEEE Computer Graphics and Applications*, *37*(5), 50–60. https://doi.org/10.1109/MCG.2017.3621223

Karimi, F., Sultana, S., Shirzadi Babakan, A., & Suthaharan, S. (2019). An enhanced support vector machine model for urban expansion prediction. *Computers, Environment and Urban Systems*, *75*(August 2018), 61–75. https://doi.org/10.1016/j.compenvurbsys.2019.01.001

Kepaptsoglou, K., Meerschaert, V., Neergaard, K., Papadimitriou, S., Rye, T., Schremser, R., & Vleugels, I. (2012). Quality Management in Mobility Management: A Scheme for Supporting Sustainable Transportation in Cities. *International Journal of Sustainable Transportation*, *6*(4), 238–256. https://doi.org/10.1080/15568318.2011.587137

Kim, J., Park, J., & Lee, W. (2018). Why do people move? Enhancing human mobility prediction using local functions based on public records and SNS data. *Plos One*, *13*, e0192698. https://doi.org/10.1371/journal.pone.0192698

Kim, Jooho, Bae, J., & Hastak, M. (2018). Emergency information diffusion on online social media during storm Cindy in U.S. *International Journal of Information Management*, *40*(January), 153–165. https://doi.org/10.1016/j.ijinfomgt.2018.02.003

Kim, Jooho, & Hastak, M. (2018). Social network analysis: Characteristics of online social networks after a disaster. *International Journal of Information Management*, *38*(1), 86–96. https://doi.org/10.1016/j.ijinfomgt.2017.08.003

Kryvasheyeu, Y., Chen, H., Obradovich, N., Moro, E., Van Hentenryck, P., Fowler, J., & Cebrian, M. (2016). Rapid assessment of disaster damage using social media activity. *Science Advances*, *2*(3), e1500779–e1500779. https://doi.org/10.1126/sciadv.1500779

Kryvasheyeu, Yury, Chen, H., Moro, E., Van Hentenryck, P., & Cebrian, M. (2015). Performance of Social Network Sensors during Hurricane Sandy. *PLOS ONE*, *10*(2), e0117288. https://doi.org/10.1371/journal.pone.0117288

Kurkcu, A., Ozbay, K., & Morgul, E. F. (2016). Evaluating the Usability of Geo-located Twitter as a Tool for Human Activity and Mobility Patterns: A Case Study for New York City. *Transportation Research Board's 95th Annual Meeting*, 1–20. Retrieved from http://ebooks.cambridge.org/ref/id/CBO9781107415324A009

Lachlan, K. A., Spence, P. R., Lin, X., & Del Greco, M. (2014). Screaming into the Wind: Examining the Volume and Content of Tweets Associated with Hurricane Sandy. *Communication Studies*, *65*(5), 500–518. https://doi.org/10.1080/10510974.2014.956941

Lagomarsino, D., Tofani, V., Segoni, S., Catani, F., & Casagli, N. (2017). A Tool for Classification and Regression Using Random Forest Methodology: Applications to Landslide Susceptibility Mapping and Soil Thickness Modeling. *Environmental Modeling & Assessment*, *22*(3), 201–214. https://doi.org/10.1007/s10666-016-9538-y

Landwehr, P. M., Wei, W., Kowalchuck, M., & Carley, K. M. (2016). Using tweets to support disaster planning, warning and response. *Safety Science*, *90*, 33–47. https://doi.org/10.1016/j.ssci.2016.04.012

Lefebvre, H. (1991). *The production of space*. Oxford, OX, UK ; Cambridge, Mass., USA : Blackwell.

Lenormand, M., Bassolas, A., & Ramasco, J. J. (2016). Systematic comparison of trip distribution laws and models. *Journal of Transport Geography*, *51*, 158–169. https://doi.org/10.1016/j.jtrangeo.2015.12.008

Leshem, G., & Ritov, Y. (2007). Traffic flow prediction using adaboost algorithm with random forests as a weak learner. *International Journal of Mathematical and Computational Sciences*, *1*(1), 193–198. https://doi.org/10.1007/s11117-011-0122-z

Li, Lifang, Zhang, Q., Tian, J., & Wang, H. (2018). Characterizing information propagation patterns in emergencies: A case study with Yiliang Earthquake. *International Journal of Information Management*, *38*(1), 34–41. https://doi.org/10.1016/j.ijinfomgt.2017.08.008

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

Lindsay, B. R. (2011). Social Media and Disasters: Current Uses, Future Options and Policy Considerations. In *Congressional Research Service Reports*. https://doi.org/R41987

Liu, B. F., Fraustino, J. D., & Jin, Y. (2015). How Disaster Information Form, Source, Type, and Prior Disaster Exposure Affect Public Outcomes: Jumping on the Social Media Bandwagon? *Journal of Applied Communication Research*, *43*(1), 44–65. https://doi.org/10.1080/00909882.2014.982685

Liu, J., Zhao, K., Khan, S., Cameron, M., & Jurdak, R. (2015). Multi-scale population and mobility estimation with geo-tagged Tweets. *2015 31st IEEE International Conference on Data Engineering Workshops*, *2015-June*, 83–86. https://doi.org/10.1109/ICDEW.2015.7129551

Liu, Y., Liu, X., Gao, S., Gong, L., Kang, C., Zhi, Y., … Shi, L. (2015). Social Sensing: A New Approach to Understanding Our Socioeconomic Environments. *Annals of the Association of American Geographers*, *105*(3), 512–530. https://doi.org/10.1080/00045608.2015.1018773

Lu, B., Charlton, M., Harris, P., & Fotheringham, A. S. (2014). Geographically weighted regression with a non-Euclidean distance metric: a case study using hedonic house price data. *International Journal of Geographical Information Science*, *28*(4), 660–681. https://doi.org/10.1080/13658816.2013.865739

Marcolin, B. L., Compeau, D. R., Munro, M. C., & Huff, S. L. (2000). Assessing User Competence: Conceptualization and Measurement. *Information Systems Research*, *11*(1), 37–60. https://doi.org/10.1287/isre.11.1.37.11782

Marston, S. A., & Smith, N. (2001). States, scales and households: limits to scale thinking? A response to Brenner. *Progress in Human Geography*, *25*(4), 615–619. https://doi.org/10.1191/030913201682688968

Martín, Y., Li, Z., & Cutter, S. L. (2017). Leveraging Twitter to gauge evacuation compliance: Spatiotemporal analysis of Hurricane Matthew. *PLOS ONE*, *12*(7), e0181701. https://doi.org/10.1371/journal.pone.0181701

Martínez-Rojas, M., Pardo-Ferreira, M. del C., & Rubio-Romero, J. C. (2018). Twitter as a tool for the management and analysis of emergency situations: A systematic literature review. *International Journal of Information Management*, *43*(August), 196–208. https://doi.org/10.1016/j.ijinfomgt.2018.07.008

May, A. D. (2013). Urban Transport and Sustainability: The Key Challenges. *International Journal of Sustainable Transportation*, *7*(3), 170–185. https://doi.org/10.1080/15568318.2013.710136

Mazhar Rathore, M., Ahmad, A., Paul, A., Hong, W.-H., & Seo, H. (2017). Advanced computing model for geosocial media using big data analytics. *Multimedia Tools and Applications*, *76*(23), 24767–24787. https://doi.org/10.1007/s11042-017-4644-7

McMaster, R. B., & Sheppard, E. (2004). Introduction: Scale and Geographic Inquiry. In R. B. McMaster & E. Sheppard (Eds.), *Scale and Geographic Inquiry* (pp. 1–22). https://doi.org/10.1002/9780470999141.ch1

Mcnally, M. G. (2007). The Four Step Model. In D. A. Hensher & K. J. Button (Eds.), *Handbook of Transport Modeling* (2nd ed., pp. 35–52). https://doi.org/https://escholarship.org/uc/item/0r75311t

McNeill, G., Bright, J., & Hale, S. A. (2017). Estimating local commuting patterns from geolocated Twitter data. *EPJ Data Science*, *6*(1), 24. https://doi.org/10.1140/epjds/s13688-017-0120-x

Miller, H. J. (2010). The data avalanche is here. Shouldn't we be digging? *Journal of Regional Science*, *50*(1), 181–201. https://doi.org/10.1111/j.1467-9787.2009.00641.x

Miller, H. J., & Goodchild, M. F. (2015). Data-driven geography. *GeoJournal*, *80*(4), 449–461. https://doi.org/10.1007/s10708-014-9602-6

Mohammadi, N., Wang, Q., & Taylor, J. E. (2016). Diffusion Dynamics of Energy Saving Practices in Large Heterogeneous Online Networks. *PLoS ONE*, *11*(10), 1–23. https://doi.org/10.1371/journal.pone.0164476

Mozolin, M., Thill, J.-C., & Lynn Usery, E. (2000). Trip distribution forecasting with multilayer perceptron neural networks: A critical evaluation. *Transportation Research Part B: Methodological*, *34*(1), 53–73. https://doi.org/10.1016/S0191-2615(99)00014-4

Muralidharan, S., Rasmussen, L., Patterson, D., & Shin, J.-H. (2011). Hope for Haiti: An analysis of Facebook and Twitter usage during the earthquake relief efforts. *Public Relations Review*, *37*(2), 175–177. https://doi.org/10.1016/j.pubrev.2011.01.010

Murthy, D., & Gross, A. J. (2017). Social media processes in disasters: Implications of emergent technology use. *Social Science Research*, *63*, 356–370. https://doi.org/10.1016/j.ssresearch.2016.09.015

Murthy, D., & Longwell, S. A. (2013). Twitter and Disasters. *Information, Communication & Society*, *16*(6), 837–855. https://doi.org/10.1080/1369118X.2012.696123

Nam, W.-H., Hayes, M. J., Svoboda, M. D., Tadesse, T., & Wilhite, D. A. (2015). Drought hazard assessment in the context of climate change for South Korea. *Agricultural Water Management*, *160*, 106–117. https://doi.org/10.1016/j.agwat.2015.06.029

National Oceanic and Atmospheric Administration (NOAA). (2013). *Service Assessment: Hurricane / Post-Tropical Cyclone Sandy , October 22 – 29 , 2012*. Retrieved from https://www.weather.gov/media/publications/assessments/Sandy13.pdf

National Weather Service. (2013). *Service Assessment: Hurricane/Post-Tropical Cyclone Sandy, October 22-29, 2012*. Retrieved from https://www.weather.gov/media/publications/assessments/Sandy13.pdf

Neppalli, V. K., Caragea, C., Squicciarini, A., Tapia, A., & Stehle, S. (2017). Sentiment analysis during Hurricane Sandy in emergency response. *International Journal of Disaster Risk Reduction*, *21*(May 2016), 213–222. https://doi.org/10.1016/j.ijdrr.2016.12.011

Nielsen, R. H. (1987). Kolmogorov's Mapping Neural Network Existence Theorem. *Proceedings of the IEEE First International Conference on Neural Networks*, 11--13. San Diego: Piscataway, NJ: IEEE.

Nimmo, B. (2019). *Measuring Traffic Manipulation on Twitter*. Retrieved from https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2019/01/Manipulating-Twitter-Traffic.pdf

Oliveira, M., & Gama, J. (2012). An overview of social network analysis. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *2*(2), 99–115. https://doi.org/10.1002/widm.1048

Ostmann, A., & Martínez Arbizu, P. (2018a). Predictive models using randomForest regression for distribution patterns of meiofauna in Icelandic waters. *Marine Biodiversity*, *48*(2), 719–735. https://doi.org/10.1007/s12526-018-0882-9

Ostmann, A., & Martínez Arbizu, P. (2018b). Predictive models using randomForest regression for distribution patterns of meiofauna in Icelandic waters. *Marine Biodiversity*, *48*(2), 719–735. https://doi.org/10.1007/s12526-018-0882-9

Park, S. J., Lim, Y. S., & Park, H. W. (2015). Comparing Twitter and You Tube networks in information diffusion: The case of the "Occupy Wall Street" movement. *Technological Forecasting and Social Change*, *95*, 208–217. https://doi.org/10.1016/j.techfore.2015.02.003

Pitombo, C. S., de Souza, A. D., & Lindner, A. (2017). Comparing decision tree algorithms to estimate intercity trip distribution. *Transportation Research Part C: Emerging Technologies*, *77*, 16–32. https://doi.org/10.1016/j.trc.2017.01.009

Pouebrahim, N., Sultana, S., Thill, J.-C., & Mohanty, S. (2018). Enhancing Trip Distribution Using Twitter Data: Comparison of Gravity and Neural Networks. *2nd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery (GeoAI'18)*. https://doi.org/10.1145/3281548.3281555

Pourebrahim, N., Sultana, S., Edwards, J., Gochanour, A., & Mohanty, S. (2019). Understanding communication dynamics on Twitter during natural disasters: A case study of Hurricane Sandy. *International Journal of Disaster Risk Reduction*, *37*(May), 101176. https://doi.org/10.1016/j.ijdrr.2019.101176

Pourebrahim, N., Sultana, S., Niakanlahiji, A., & Thill, J.-C. (2019). Trip distribution modeling with Twitter data. *Computers, Environment and Urban Systems*, *77*(June 2019), 101354. https://doi.org/10.1016/j.compenvurbsys.2019.101354

Pred, A. (1984). Place as Historically Contingent Process : Structuration and the Time-Geography of Becoming Places. *Annals of the Association of American Geographers*, *74*(2), 279–297.

Qadir, J., Ali, A., ur Rasool, R., Zwitter, A., Sathiaseelan, A., & Crowcroft, J. (2016). Crisis analytics: big data-driven crisis response. *Journal of International Humanitarian Action*, *1*(1), 12. https://doi.org/10.1186/s41018-016-0013-9

Ragini, J. Rexiline, Anand, P. M. R., & Bhaskar, V. (2018). Big data analytics for disaster response and recovery through sentiment analysis. *International Journal of Information Management*, *42*(May), 13–24. https://doi.org/10.1016/j.ijinfomgt.2018.05.004

Ragini, J.R, Rubesh Anand, P. M., & Bhaskar, V. (2018). Mining crisis information: A strategic approach for detection of people at risk through social media analysis. *International Journal of Disaster Risk Reduction*, *27*(August 2017), 556–566. https://doi.org/10.1016/j.ijdrr.2017.12.002

Rao, A., Spasojevic, N., Li, Z., & Dsouza, T. (2015). Klout score: Measuring influence across multiple social networks. *2015 IEEE International Conference on Big Data (Big Data)*, (October 2015), 2282–2289. https://doi.org/10.1109/BigData.2015.7364017

Rashidi, T. H., & Mohammadian, A. (2011). Household travel attributes transferability analysis: Application of a hierarchical rule based approach. *Transportation*, *38*(4), 697–714. https://doi.org/10.1007/s11116-011-9339-8

Rasouli, S., & Timmermans, H. J. P. (2014). Using ensembles of decision trees to predict transport mode choice decisions: Effects on predictive success and uncertainty estimates. *EJTIR Issue*, *14*(4), 412–424.

Roy, J. R., & Thill, J. C. (2003). Spatial interaction modelling. *Papers in Regional Science*, *83*(1), 339–361. https://doi.org/10.1007/s10110-003-0189-4

Rumelhart, D. E., & Mcclelland, J. L. (1986). *Parallel distributed processing: explorations in the microstructure of cognition. Volume 1. Foundations*. Retrieved from https://www.researchgate.net/publication/200033859

Saleem, H. M., Xu, Y., & Ruths, D. (2014a). Effects of Disaster Characteristics on Twitter Event Signature. *Procedia Engineering*, *78*, 165–172. https://doi.org/10.1016/j.proeng.2014.07.053

Saleem, H. M., Xu, Y., & Ruths, D. (2014b). Novel Situational Information in Mass Emergencies: What does Twitter Provide? *Procedia Engineering*, *78*, 155–164. https://doi.org/10.1016/j.proeng.2014.07.052

Salton, G., & Buckley, C. (1988). Text Retrieval. *Information Processing & Management*, *24*(5), 513–523.

Sapountzi, A., & Psannis, K. E. (2018). Social networking data analysis tools & challenges. *Future Generation Computer Systems*, *86*, 893–913. https://doi.org/10.1016/j.future.2016.10.019

Sekhar, C. R., Minal, & Madhu, E. (2016). Mode Choice Analysis Using Random Forrest Decision Trees. *Transportation Research Procedia*, *17*(December 2014), 644–652. https://doi.org/10.1016/j.trpro.2016.11.119

Shaw, S. L., Tsou, M. H., & Ye, X. (2016). Editorial: human dynamics in the mobile and big data era. *International Journal of Geographical Information Science*, *30*(9), 1687–1693. https://doi.org/10.1080/13658816.2016.1164317

Shelton, T., Poorthuis, A., Graham, M., & Zook, M. (2014). Mapping the data shadows of Hurricane Sandy: Uncovering the sociospatial dimensions of "big data." *Geoforum*, *52*, 167–179. https://doi.org/10.1016/j.geoforum.2014.01.006

Shirzadi Babakan, A., Alimohammadi, A., & Taleai, M. (2015). An agent-based evaluation of impacts of transport developments on the modal shift in Tehran, Iran. *Journal of Development Effectiveness*, *7*(2), 230–251. https://doi.org/10.1080/19439342.2014.994656

Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. *Proceedings Ofthe Workshop on Interactive Language Learning, Visualization, and Interfaces*, 63–70. https://doi.org/10.3115/v1/w14-3110

Simini, F., González, M. C., Maritan, A., & Barabási, A.-L. (2012). A universal model for mobility and migration patterns. *Nature*, *484*(7392), 96–100. https://doi.org/10.1038/nature10856

SimplyAnalytics. (2015). Retrieved from http://simplyanalytics.com/

Sit, M. A., Koylu, C., & Demir, I. (2019). Identifying disaster-related tweets and their semantic, spatial and temporal context using deep learning, natural language processing and spatial analysis: a case study of Hurricane Irma. *International Journal of Digital Earth*, (January). https://doi.org/10.1080/17538947.2018.1563219

Soja, E. (1999). Thirdspace: Expanding the scope of the geographical imagination. In D. Massey, O. Allen, & P. Sarre (Eds.), *Human geography today* (p. 260). Cambridge: Polity Press.

Sorensen, J. H., Sorensen, B. V., Smith, A., & Williams, Z. (2009). *Results of An Investigation of the Effectiveness of Using Reverse Telephone Emergency Warning Systems in the October 2007 San Diego Wildfires*. https://doi.org/ORNL/TM-2009/154

Spence, P. R., Lachlan, K. A., Lin, X., & del Greco, M. (2015). Variability in Twitter Content Across the Stages of a Natural Disaster: Implications for Crisis Communication. *Communication Quarterly*, *63*(2), 171–186. https://doi.org/10.1080/01463373.2015.1012219

Srisaeng, P., & Baxter, G. (2017). Modelling Australia's outbound passenger air travel demand using an artificial neural network approach. *International Journal For Traffic And Transport Engineering*, *7*(4), 406–423. https://doi.org/10.7708/ijtte.2017.7(4).01

Stathopoulos, A., Dimitriou, L., & Tsekeris, T. (2008). Fuzzy modeling approach for combined forecasting of urban traffic flow. *Computer-Aided Civil and Infrastructure Engineering*, *23*(7), 521–535. https://doi.org/10.1111/j.1467-8667.2008.00558.x

Steiger, E., de Albuquerque, J. P., & Zipf, A. (2015). An Advanced Systematic Literature Review on Spatiotemporal Analyses of Twitter Data. *Transactions in GIS*, *19*(6), 809–834. https://doi.org/10.1111/tgis.12132

Stewart, M. C., & Gail Wilson, B. (2016). The dynamic role of social media during Hurricane #Sandy: An introduction of the STREMII model to weather the storm of the crisis lifecycle. *Computers in Human Behavior*, *54*, 639–646. https://doi.org/10.1016/j.chb.2015.07.009

Sui, D., & Goodchild, M. (2011). The convergence of GIS and social media: Challenges for GIScience. *International Journal of Geographical Information Science*, *25*(11), 1737–1748. https://doi.org/10.1080/13658816.2011.604636

Sulaiman, S., Shamsuddin, S. M., Abraham, A., & Sulaiman, S. (2011). Intelligent web caching using machine learning methods. *Neural Network World*, *21*(5), 429–452. https://doi.org/10.14311/NNW.2011.21.025

Sultana, S., Pourebrahim, N., & Kim, H. (2018). Household Energy Expenditures in North Carolina: A Geographically Weighted Regression Approach. *Sustainability*, *10*(5), 1511. https://doi.org/10.3390/su10051511

Sultana, S., & Weber, J. (2007). Journey-to-Work Patterns in the Age of Sprawl: Evidence from Two Midsize Southern Metropolitan Areas*. *The Professional Geographer*, *59*(2), 193–208. https://doi.org/10.1111/j.1467-9272.2007.00607.x

Sultana, S., & Weber, J. (2014). The Nature of Urban Growth and the Commuting Transition: Endless Sprawl or a Growth Wave? *Urban Studies*, *51*(3), 544–576. https://doi.org/10.1177/0042098013498284

Sutton, J., Palen, L., & Shklovski, I. (2008). Backchannels on the Front Lines: Emergent Uses of Social Media in the 2007 Southern California Wildfires. *Proceedings of the 5 Th International ISCRAM Conference*, (May), 624–632.

Swyngedouw, E. (2004). Scaled Geographies: Nature, Place, and the Politics of Scale. In E. Sheppard & R. B. McMaster (Eds.), *Scale and Geographic Inquiry* (pp. 129–153). https://doi.org/10.1002/9780470999141.ch7

Tang, Z., Zhang, L., Xu, F., & Vo, H. (2015). Examining the role of social media in California's drought risk management in 2014. *Natural Hazards*, *79*(1), 171–193. https://doi.org/10.1007/s11069-015-1835-2

Taylor, P. J. (1982). A Materialist Framework for Political Geography. *Transactions of the Institute of British Geographers*, *7*(1), 15. https://doi.org/10.2307/621909

Thill, J.-C., & Wheeler, A. (2000a). Knowledge discovery and induction of decision trees in spatial decision problems. In A. Reggiani (Ed.), *Spatial economic science: New frontiers in theory and methodology* (pp. 188–205). https://doi.org/10.1007/978-3-642-59787-9_10

Thill, J.-C., & Wheeler, A. (2000b). Tree Induction of Spatial Choice Behavior. *Transportation Research Record: Journal of the Transportation Research Board*, *1719*(1), 250–258. https://doi.org/10.3141/1719-33

Thrift, N. (2008). *Non-representational theory: Space, politics, affect*. Milton Park, Abingdon: Oxon: Routledge.

Tillema, F., van Zuilekom, K. M., & van Maarseveen, M. F. A. M. (2006). Comparison of Neural Networks and Gravity Models in Trip Distribution. *Computer-Aided Civil and Infrastructure Engineering*, *21*(2), 104–119. https://doi.org/10.1111/j.1467-8667.2005.00421.x

Tsou, M.-H., Zhang, H., & Jung, C.-T. (2017). Identifying Data Noises, User Biases, and System Errors in Geo-tagged Twitter Messages (Tweets). *Nan*, *nan*(nan), nan. https://doi.org/None

Tsou, M. H. (2015). Research challenges and opportunities in mapping social media and Big Data. *Cartography and Geographic Information Science*, *42*(1), S70–S74. https://doi.org/10.1080/15230406.2015.1059251

Tsou, M., Jung, C., Allen, C., Yang, J., Han, S. Y., Spitzberg, B. H., & Dozier, J. (2017). *Building a Real-Time Geo-Targeted Event Observation (Geo) Viewer for Disaster Management and Situation Awareness*. https://doi.org/10.1007/978-3-319-57336-6_7

Tuan, Y. F. (1977). *Space and place: The perspective of experience*. Minneapolis, MN: University of Minnesota Press.

Tyshchuk, Y., Hui, C., Grabowski, M., & Wallace, W. A. (2012). Social Media and Warning Response Impacts in Extreme Events: Results from a Naturally Occurring Experiment. *2012 45th Hawaii International Conference on System Sciences*, 818–827. https://doi.org/10.1109/HICSS.2012.536

Ukkusuri, S. V., Zhan, X., Sadri, A. M., & Ye, Q. (2014). Use of Social Media Data to Explore Crisis Informatics. *Transportation Research Record: Journal of the Transportation Research Board*, *2459*(1), 110–118. https://doi.org/10.3141/2459-13

Vlahogianni, E. I., Karlaftis, M. G., & Golias, J. C. (2007). Spatio-temporal short-term urban traffic volume forecasting using genetically optimized modular networks. *Computer-Aided Civil and Infrastructure Engineering*, *22*(5), 317–325. https://doi.org/10.1111/j.1467-8667.2007.00488.x

Vo, B.-K. H., & Collier, N. (2013). Twitter Emotion Analysis in Earthquake Situations. *International Journal of Computational Linguistics and Applications*, *4*(1), 159–173.

Wang, Z., Lam, N. S. N., Obradovich, N., & Ye, X. (2019). Are vulnerable communities digitally left behind in social responses to natural disasters? An evidence from Hurricane Sandy with Twitter data. *Applied Geography*, *108*(April), 1–8. https://doi.org/10.1016/j.apgeog.2019.05.001

Wang, Z., & Ye, X. (2018). Space, time, and situational awareness in natural hazards: a case study of Hurricane Sandy with social media data. *Cartography and Geographic Information Science*, *00*(00), 1–13. https://doi.org/10.1080/15230406.2018.1483740

Wang, Z., Ye, X., & Tsou, M. H. (2016). Spatial, temporal, and content analysis of Twitter for wildfire hazards. *Natural Hazards*, *83*(1), 523–540. https://doi.org/10.1007/s11069-016-2329-6

Waters, R. D., & Williams, J. M. (2011). Squawking, tweeting, cooing, and hooting: analyzing the communication patterns of government agencies on Twitter. *Journal of Public Affairs*, *11*(4), 353–363. https://doi.org/10.1002/pa.385

Westerman, D., Spence, P. R., & Van Der Heide, B. (2012). A social network as information: The effect of system generated reports of connectedness on credibility on Twitter. *Computers in Human Behavior*, *28*(1), 199–206. https://doi.org/10.1016/j.chb.2011.09.001

Williams, H. T. P., McMurray, J. R., Kurz, T., & Hugo Lambert, F. (2015). Network analysis reveals open forums and echo chambers in social media discussions of climate change. *Global Environmental Change*, *32*, 126–138. https://doi.org/10.1016/j.gloenvcha.2015.03.006

Wilson, A. (1970). *Entropy in Urban and Regional Modelling*. London: Routledge.

Wilson, A. G. (1998). Land-use / Transport Interaction Models Past and Future. *Journal of Transport Economics and Policy*, *32*(1), 3–26.

Yaldi, G., Taylor, M. A. P., & Yue, W. L. (2009). Improving artificial neural network performance in calibrating doubly-constrained work trip distribution by using a simple data normalization and linear activation function. *32nd Australasian Transport Research Forum, ATRF 2009*, (February 2015).

Yaldi, G., Taylor, M. A. P., & Yue, W. L. (2011). Forecasting origin-destination matrices by using neural network approach: A comparison of testing performance between back propagation, variable learning rate and levenberg-marquardt algorithms. *Australasian Transport Research Forum 2011*, (September), 1–15. Retrieved from https://pdfs.semanticscholar.org/4c3f/8281bf07e838294cc5209f6da6cf37c19d4a.pdf

Yang, F., Jin, P. J., Cheng, Y., Zhang, J., & Ran, B. (2015). Origin-Destination Estimation for Non-Commuting Trips Using Location-Based Social Networking Data. *International Journal of Sustainable Transportation*, *9*(8), 551–564. https://doi.org/10.1080/15568318.2013.826312

Yang, J.-A., Tsou, M.-H., Jung, C.-T., Allen, C., Spitzberg, B. H., Gawron, J. M., & Han, S.-Y. (2016). Social media analytics and research testbed (SMART): Exploring spatiotemporal patterns of human dynamics with geo-targeted social media messages. *Big Data & Society*, *3*(1), 205395171665291. https://doi.org/10.1177/2053951716652914

Yang, X. (2015). *Exploring the World with Volunteered Geographic Information: Space, Place and People* (The Ohio State University). Retrieved from https://etd.ohiolink.edu/!etd.send_file?accession=osu1429609791&disposition=inline

Yang, X., Ye, X., & Sui, D. Z. (2016). We know where you are: In space and place - Enriching the geographical context through social media. *International Journal of Applied Geospatial Research*, *7*(2), 61–75. https://doi.org/10.4018/IJAGR.2016040105

Yang, Y. (2013). *Understanding Human Mobility Patterns from Digital Traces* (Massachusetts Institute of Technology). Retrieved from http://hdl.handle.net/1721.1/82863

Yang, Y., Herrera, C., Eagle, N., & González, M. C. (2015). Limits of Predictability in Commuting Flows in the Absence of Data for Calibration. *Scientific Reports*, *4*(1), 5662. https://doi.org/10.1038/srep05662

Yuan, F., & Liu, R. (2018). Feasibility study of using crowdsourcing to identify critical affected areas for rapid damage assessment: Hurricane Matthew case study. *International Journal of Disaster Risk Reduction*, *28*(February), 758–767. https://doi.org/10.1016/j.ijdrr.2018.02.003

Zhang, Y., & Haghani, A. (2015). A gradient boosting method to improve travel time prediction. *Transportation Research Part C: Emerging Technologies*, *58*, 308–324. https://doi.org/10.1016/j.trc.2015.02.019

Zipf, G. K. (1946). The P 1 P 2 /D Hypothesis: On the Intercity Movement of Persons. In *Source: American Sociological Review* (Vol. 11).

Zook, M. A., & Graham, M. (2007a). Mapping DigiPlace: Geocoded Internet data and the representation of place. *Environment and Planning B: Planning and Design*, *34*(3), 466–482. https://doi.org/10.1068/b3311

Zook, M. A., & Graham, M. (2007b). The creative reconstruction of the Internet: Google and the privatization of cyberspace and DigiPlace. *Geoforum*, *38*(6), 1322–1343. https://doi.org/10.1016/j.geoforum.2007.05.004

Zou, L., Lam, N. S. N., Cai, H., & Qiang, Y. (2018). Mining Twitter Data for Improved Understanding of Disaster Resilience. *Annals of the American Association of Geographers*, *4452*, 1–20. https://doi.org/10.1080/24694452.2017.1421897