Tweet diffusion over space and time: A comparison of a winter-weather and tornado-outbreak case study

A Thesis Presented for the

Master of Science

Degree

The University of Tennessee, Knoxville

Alyssa Margaret Cannistraci

December 2020

Copyright © 2020 by Alyssa Margaret Cannistraci

All rights reserved

DEDICATION

My thesis is dedicated to my parents, neither of whom attended a four-year university or graduate school. Though the idea of higher education was unfamiliar, they supported my dreams unconditionally. They taught me the value of perseverance, integrity, and empathy. My parents constantly reminded me of the exciting world beyond our hometown, and encouraged me to seek opportunities. They molded me into the woman I am today. My parents are amazing people and I hope my thesis makes them proud.

ACKNOWLEDGMENTS

I would first like to thank my parents and sister. Their encouragement motivated me to persevere through challenges, and showed me my scholarly potential. Friends listened to ideas, provided constructive criticism, and offered advice through all stages of my thesis. I am thankful to receive funding from the University of Tennessee, Knoxville to pursue my graduate work. I would also like to thank my advisor, Dr. Kelsey Ellis, for selecting me for the master's program and guiding me through my research. Her mentorship helped me improve my writing style and build confidence. Additionally, I would like to thank Dr. Shih-Lung Shaw and Dr. Qiusheng Wu for serving on my thesis committee. Both offered further direction on my project, and provided knowledge in content-areas for which I am less familiar. Dr. Yingjie Hu (University of Buffalo) acquired tweets for me to analyze, and I am appreciative of his time. On a lighter note, I would also like to thank Dunkin' (and their excellent rewards program) for providing the caffeine needed to complete this degree. And finally, thank you to my beloved hamster Trufflez for making me smile.

ABSTRACT

Social media allows people to receive, engage in, and share weather information. Users of the social media platform Twitter actively share weather content via tweets, which researchers can acquire through an Application Programming Interface (API). APIs return tweet content, as well as temporal and spatial characteristics (latitude and longitude coordinates). Tweets can then be mapped and studied spatiotemporally through Geographic Information System (GIS) software. For this work, I compared how tweets spread ("diffuse") over space and time during two natural hazard events in the United States. The first case study is a winter weather event that The Weather Channel named "Winter Storm Gia" (WSG). WSG occurred from 11 to 14 January 2019 and impacted the Rockies, Mid-West, Ohio River Valley, and Mid-Atlantic. The second case study is a tornado outbreak in the Southeast that occurred in early March 2019, and resulted in 80 tornadoes ("Storm Prediction Center Storm Reports for 3/3/19," 2019). I adapted the methods used by Lee, Lay, Chin, Chi, and Hsueh (2014); specifically, I used the Nearest Neighbor Index (NNI) and regression analysis to study tweet diffusion. I also examined the tweet locations using the ArcGIS "Directional Distribution (Standard Deviational Ellipse)" tool. For both events, tweets peaked on the event day, but for the tornado outbreak, there was also a rebound in tweets after the event that was not present after WSG. Tornado-outbreak tweets were also less concentrated in cities than was the case in WSG. This suggests population dynamics (distance to individual tornadoes) affects winter weather (tornado-outbreak) tweet diffusion. Lastly, I found regression analysis of NNI values not to be a suitable method for analyzing tweet diffusion of weather events. The results were limited by a small sample size of tornado-outbreak tweets. Understanding how weather

information spreads on Twitter is important for individuals who communicate information to communities. This includes National Weather Service (NWS) forecasters, emergency managers, media personnel, and governmental officials.

TABLE OF CONTENTS

INTRODUCTION 1

BACKGROUND 3

2.1 Twitter data in academic research 3

2.2 Twitter Application Programing Interfaces (APIs) 4

2.3 Twitter within the weather community 6

2.4 Causes and types of diffusion 7

2.5 Spatiotemporal diffusion barriers 9

2.6 Quantitatively studying spatial diffusion 11

DATA AND METHODS 14

3.1 Case studies 14

3.2 Data 16

3.2.a Meteorological data 16

3.2.b Twitter data 16

3.3 Filtering tweets using different approaches 17

3.3.a Filtering tweets by time and location 17

3.3.b Filtering tweets by relevancy 23

3.4 Nearest Neighbor Index (NNI) 29

3.5 NNI and regression analysis 30

3.6 Standard Deviational Ellipse (SDE) 32

RESULTS 34

4.1 Daily tweet counts over case study periods 34

4.2 Spatiotemporal diffusion regression analysis 36

4.3 Standard Deviational Ellipse (SDE) analysis 38

DISCUSSION 50

5.1 WSG diffusion over space and time 50

5.2 Tornado-outbreak tweets over space and time 51

5.3 Tweet diffusion barriers during hazardous weather events 54

5.4 Uncertainties using regression to analyze spatiotemporal diffusion events 54

5.5 Limitations 56

CONCLUSION 58
REFERENCES 62
VITA 70

LIST OF TABLES

Table 1: API keywords 19
Table 2: Case study date ranges 22
Table 3: Properties of the SDE mean centers for WSG tweets 41
Table 4: Properties of SDE area for WSG tweets 43
Table 5: Properties of SDE mean centers for the tornado-outbreak tweets 45
Table 6: Properties of SDE area for the tornado-outbreak tweets 48

LIST OF FIGURES

Figure 1: API sample region 18
Figure 2: Irrelevant WSG tweets 25
Figure 3: Irrelevant tornado-outbreak tweets (first filter pass) 27
Figure 4: Irrelevant tornado-outbreak tweets (second filter pass) 28
Figure 5: Daily tweet counts 35
Figure 6: WSG tweets on two different days 35
Figure 7: Tornado outbreak tweets on two different days 37
Figure 8: WSG regression analysis 37
Figure 9: Tornado-outbreak regression analysis 39
Figure 10: SDE mean center location reference axes 39
Figure 11: Daily mean center locations for WSG tweets 42
Figure 12: WSG SDE maps 44
Figure 13: Daily mean center locations for tornado-outbreak tweets 46
Figure 14: Tornado-outbreak SDE maps 49

CHAPTER ONE

INTRODUCTION

Technologic innovations result in improved weather forecasting and communication, allowing for the public to access and share continuous weather information (Morss et al., 2017). However, lives are still lost to hazardous weather events, even with advanced forecasting and communication techniques available. The infamous April 2011 tornado outbreak demonstrates this discontinuity, as all 300 fatalities from this event were located within counties warned for severe weather by the National Weather Service (NWS) (Hayes, 2012). There appears to be some barriers to risk communication during hazardous weather, as effective weather communication remains a challenge. Ripberger et al. (2014) found the public listens to some, but not all tornado-related information, communicated by the NWS. Communication issues are likely seen across many natural hazard events, and these barriers are not just specific to weather events.

Social media is a relatively new technology that can be used to help spread information before, during, and after weather events (Houston et al., 2015). There are several social media platforms, one of them being Twitter. Twitter is consistently ranked a top social media platform annually across usage surveys (Perrin & Anderson, 2019), and is used during hazardous weather events (Cooper et al., 2015a, 2015b; Ripberger et al., 2014). The demographic of Twitter users largely includes young (18 to 24 years old), educated (college-level) individuals who live in urban locations (Perrin & Anderson, 2019). Twitter users interact through tweets that include written or multimedia commentary, formatted in 280 characters or less (“Counting characters,” 2020). Tweets can include words, pictures, videos, or links to other online content (“About

different types of Tweets," 2020). It is not uncommon to see tweets that discuss or debate current events (Palen, Vieweg, Liu, & Hughes, 2009; Qu, Wu, & Wang, 2009).

During a hazardous weather event, Twitter may be used to spread information from forecasters to the public. Ripberger et al. (2014) analyzed tweets during the 2012 summer season and found the presence of tornado watches and warnings triggered heightened tornado-risk awareness in Twitter users. Ripberger et al. (2014) claimed hazardous weather messaging must be "receive[d], [paid] attention to, and [understood]" to resonate with the general public. The public's receptiveness to weather risks may even contain a temporal element. According to Mason et al. (2018), Tennesseans believed their likelihood of receiving information about tornado risks is lower at night. Additionally, approximately 25% of Tennesseans used social media to receive weather updates (Mason et al., 2018). Thus, social media's use during weather events should continue to be explored.

In my thesis, I compared how the diffusion of tweets varied across two natural hazard case studies in the United States. Winter Storm Gia (WSG), a winter event given this name by The Weather Channel, is my first case study. WSG developed in early to mid-January 2019. It caused flight delays and power outages across the Rockies, Mid-West, Ohio River Valley, and Mid-Atlantic from 11 to 14 January 2019. A tornado outbreak, located in the Southeast, is my second case study. It caused 80 tornadoes, 23 fatalities, and 104 injuries in early March 2019 ("Storm Prediction Center Storm Reports for 3/3/19," 2019). The research question I aim to answer about these events is: How does the diffusion of tweets vary between a multi-day, synoptic-level winter storm and a regional, one-day tornado outbreak.

CHAPTER TWO

BACKGROUND

2.1 Twitter data in academic research

Twitter data are researched across many academic disciplines. Atmospheric science (Ripberger et al., 2014), emergency management (Bohannon, 2016), epidemiology (Bahk et al., 2016), marketing (Paniagua & Sapena, 2014), risk perception (Hoekstra et al., 2011), and data science (Hu, 2018) are among the disciplines with peer-reviewed literature involving Twitter. Tweets can reveal socioeconomic health patterns (Nguyen et al., 2017), spatial movement of individuals during a hurricane evacuation (Martín, Li, & Cutter, 2017), skepticism regarding vaccines (Bahk et al., 2016), and hysteria over H1N1 (Chew & Eysenbach, 2010) or influenza outbreaks (Stilo, Velardi, Tozzi, & Gesualdo, 2014). Twitter, therefore, offers exciting research opportunities.

Since many social media users write in the first-person, tweets can provide insight into one's life experience (Alexander, 2014; Ukkusuri, Zhan, Sadri, & Ye, 2014). For example, tweets posted during a severe weather event can reveal what people are thinking, feeling, seeing, and experiencing before, during, and after the event (Alexander, 2014; Houston et al., 2015; Qu et al., 2009; Ripberger et al., 2014). Scholars also find Twitter an interesting data source for research because tweets are posted in "quasi-real time" (Alexander, 2014; Houston et al., 2015; Martín et al., 2017). Thus, researchers can analyze first-hand accounts as weather events are happening in near real-time.

2.2 Twitter Application Programing Interfaces (APIs)

Application Programming Interfaces (APIs) are required to access Twitter data and associated metadata (Zornoza, 2019). APIs act as conduits and connect researchers to databases where information is stored (Hoffman, 2018). Individuals must adhere to specific Twitter protocols to gain API access. A special type of login, known as a Developer Account, must be created and authorized (Zornoza, 2019). Developers are assigned an application with four passwords that together allow API access ("Access tokens," 2019): Access token; Access token secret; Consumer key; and Consumer key secret. When these four passwords are incorporated into coding scripts with additional conditional logic, tweets specific to the researcher's questions can be obtained. Twitter offers various official APIs that have different capabilities, shortcomings, data limits, and prices (Kim, Nordgren, & Emery, 2020). Some APIs are free (e.g., Streaming and Searching) while others cost money (e.g., Firehose and PowerTrack).

Streaming and Searching are Twitter's two free APIs. They differ in the way tweets are returned to the researcher. The Streaming API is a "push" system because tweets are delivered (pushed) in real-time, as they are posted to Twitter (Kim et al., 2020). Meanwhile, the Searching API is a "pull" system because tweets are queried (pulled) from storage upon request (Kim et al., 2020). The Searching API is subjected to a seven-day window ("Search Tweets Standard search API," 2020), meaning only tweets from the prior week are searchable. Searching the history of a specific Twitter user's profile is one work around to this limited time window. The Streaming API is not limited to seven days; however, due to its "on-demand" nature and inability to search into the past, the Streaming API is most appropriate when researchers have

adequately planned and strategized parameters to enter into the API (Kim et al., 2020). The Searching API is better suited for more spontaneous events where the researcher may want to look back in time for tweets.

The Searching and Streaming APIs are not without limitations. Concerns begin with the API tweet sampling itself (Wang, Callan, & Zheng, 2015). The Streaming API returns up to 1% of tweets shared publicly that match queried criteria (Morstatter, Pfeffer, & Liu, 2014). It is important to note that samples returned by APIs are inconsistent, and probably not truly indicative of all possible tweets that meet certain criteria. Kim et al. (2020), for example, noticed vast differences in sampled tweet quality throughout the duration of their study. Additionally, confusion results from Twitter lacking adequate API documentation. Many researchers accept the APIs as "black boxes," and do not fully understand how they work (Morstatter et al., 2014; Wang et al., 2015). I found documentation lacking, vague, and too advanced for the layperson. Kim et al. (2020) agreed, writing, "to date, no clear guidance exists about the advantages and limitations of each API" and "data vendors provide insufficient technical guidance [for researchers] to make informed or transparent decisions" (page 1). Biases—via Twitter's API sampling algorithm (Morstatter et al., 2014) as well as socioeconomic status, geographic location, age, and demographics (Malik, Lamba, & Nakos, 2015)—are also reported. Nevertheless, despite known limitations, many researchers continue to use Twitter's free APIs over more expensive options.

Firehose and PowerTrack are two of Twitter's paid API options. Though arguably more powerful than the free APIs, sources suggest that Firehose and PowerTrack are quite expensive (Morstatter et al., 2014; Morstatter, Pfeffer, Liu, & Carley, 2013). Both claim access to every

public tweet since Twitter began in 2006 (Kim et al., 2020); thus, a large advantage over the free APIs is the sheer number of tweets available. The main difference between Firehose and PowerTrack is the way the data are processed. PowerTrack improves tweet accessibility because it allows greater control, flexibility, and filtering techniques than Firehose (Kim et al., 2020). Because Firehose and PowerTrack can easily return one million tweets or more per search, without access to computer scientists or individuals specializing in machine learning, data processing and cleaning can quickly become overwhelming.

2.3 Twitter within the weather community

The weather community has begun using Twitter data to observe how the general public experiences weather. Houston et al. (2015) detailed how people use social media before, during, and after disasters. Before a disastrous weather event, governmental leaders or local news sources might tweet informative or educational content. The general public can engage with these posts and learn how to remain safe (Houston et al., 2015). During a weather event, the general public may share live reports or alert loved ones of safety via Twitter (Houston et al., 2015). Tweets after weather events may lead to disaster-related volunteering or fundraising opportunities (Houston et al., 2015). Houston et al. (2015) also noted that socioeconomics may be a conversation topic after disasters, as people become more aware of the affected communities and recognize any social injustices.

Major weather events are sometimes studied using Twitter data. Ukkusuri et al. (2014) examined the content of tweets from the 2013 tornado in Moore, Oklahoma. Many tweets exhibited themes similar to those discussed by Houston et al. (2015); for example, there were tweets reporting tornado damage locations, identifying needs of local residents, and

coordinating fundraising efforts (Ukkusuri et al. 2014). Bohannon (2016) and Kryvasheyeu et al. (2016) analyzed Twitter data from Superstorm Sandy (2012). Their results showed that east coast emergency management agencies used Twitter to coordinate storm relief efforts (Bohannon, 2016; Kryvasheyeu et al., 2016). Emergency managers tracked Sandy-related tweets and immediately allocated resources to locations of Sandy tweet hotspots (Bohannon, 2016). Martín et al. (2017) studied Twitter activity in coastal South Carolina before and after Hurricane Matthew (2016) warnings were issued. Based upon locations of Twitter usage, Martín et al. (2017) extrapolated the percentage of community members who evacuated due to the hurricane, and their result nearly matched the evacuation rate often seen in post-hurricane surveys (Martín et al., 2017).

2.4 Causes and types of diffusion

Agents are anything that causes, propagates, or leads to diffusion—the spread of events or information over space and time. Diffusion and agents were not clearly defined in any particular study; rather, these definitions came from a synthesis of ideas from various studies (Altay & Pal, 2014; Le Texier & Caruso, 2017; Schramm, Trainor, Shanker, & Hu, 2010). The term agent is also commonly found in epidemiology literature. The Centers for Disease Control and Prevention defines agents as factors that lead to disease ("Principles of Epidemiology: Glossary," 2014). Twitter is an agent because it leads to the spread of information. Tweets are also examples of agents because they spread ideas.

Gould (1969) described four types of diffusion: hierarchical, expansion, contagious, and relocation diffusion. Spatiotemporal events likely exhibit at least one type of diffusion. However, since diffusion is largely a byproduct of human dynamics, diffusion type classification

may change over time. Spatiotemporal events may even simultaneously display multiple diffusion types (Gould, 1969).

Hierarchical diffusion occurs when information is vetted before spreading (Gould, 1969), and is similar to the idea of "gate keeping" in communication studies (Welbers & Opgenhaffen, 2018). Certain authoritative individuals, known as "gate keepers," decide and selectively filter what information is shared (Welbers & Opgenhaffen, 2018). As Gould (1969) explained, a hierarchical structure is evident as the idea or object spreads. Few people are at the top of the hierarchy. The number of people interacting and exposed to the agent increases over time.

Expansion diffusion occurs when diffusion increases with time (Gould, 1969). "Gate keeping," bureaucratic levels, and regulations are not present in expansion diffusion, which differentiates it from hierarchical diffusion. Certain individuals can, however, affect the rate of expansion diffusion (Welbers & Opgenhaffen, 2018). Malcolm Gladwell, author of *The tipping point: How little things can make a big difference*, explained this idea in his "Law of the Few" (Gladwell, 2002). Gladwell (2002) claimed these individuals—called "Connectors," "Mavens," and "Salesmen"—effectively join, acquire, and sell ideas. Connectors, mavens, and salesmen likely increase the speed of expansion diffusion.

Contagious diffusion spreads by direct contact (Gould, 1969). Direct contact takes multiple forms. Physical distance is one form of direct contact, as people in close proximity tend to communicate and share ideas freely. Virtual space, defined by Shaw and Sui (2018) as "cyberspace, telecommunication space, and social space," provides another means for direct contact (Ma, 2011). Text messages and email communication are examples of direct contact via

the virtual space, as recipients receive information directly from senders. Messages sent through social media, such as tweets, are also examples of contagious diffusion that occur in the virtual space (Welbers & Opgenhaffen, 2018).

Finally, relocation diffusion occurs when information or ideas are transported from one location to another (Gould, 1969). People often cause relocation diffusion; ideas, trends, and societal beliefs “relocate” as people physically move from one location to another (Gould, 1969). Relocation diffusion in the virtual space happens when people share content across social media platforms. Each social media platform has different user demographics (Perrin & Anderson, 2019), and different networks of people. Therefore, content shared across social media (say, Instagram to Twitter) transports (“relocates”) information.

2.5 Spatiotemporal diffusion barriers

Various factors can affect how diffusing agents spread spatiotemporally. Gould (1969) describes four different classifications of diffusion “barriers:” Technological, cultural, psychological, and physical barriers. Each barrier slows the spread of diffusing agents (Gould, 1969).

Technological barriers occur when electronic resources are lacking, unavailable, inadequate, or faulty (Gould, 1969). Some aspect of technology prevents one from accessing or sharing digital information. Users access Twitter via cellular data or WIFI connections, both of which may be weak and unreliable in rural communities. Thirteen percent of Twitter users live in rural locations (Perrin & Anderson, 2019). Technological barriers can therefore isolate some

Twitter users. Studies relying on Twitter may have skewed or biased data if voices are not completely inclusive.

Cultural barriers are socially constructed sociological groups (Gould, 1969). Examples include belief system, education level, dialect, family heritage, language spoken, linguistics, morals, opinions, political affiliation, religious practice, sexual orientation, social class, and tax bracket (Gould, 1969). These groups, though socially constructed, can lead to tension. Critical Race (Vaught & Castagno, 2008) and Marxist (Wolff & Zalta, 2017) theories describe some conflicts. Cultural barriers can prevent diffusion agents from spreading if people are not willing to cross groups (Gould, 1969), thus, impacting spatiotemporal diffusion. Cultural barriers may even affect who uses social media. A lack of user diversity may exist since Twitter skews towards younger and more educated individuals (Perrin & Anderson, 2019). Some people may be less-receptive to Twitter or have preconceived beliefs and opinions towards the platform.

Unlike cultural barriers, which are sociologically driven, psychological barriers are perceptions of the mind (Gould, 1969). One's own internal dialog influences actions that can affect how diffusion agents spread. Gould (1969), Hägerstrand (1967), and Gladwell (2002) discuss observed patterns of individual behavior. The terminology used to describe behaviors include "Inventors," "Early Adopters," "Early Majority," "Late Majority," and "Laggards" (Gladwell, 2002; Gould, 1969). When the terms are considered in this order, they make an ordinal scale that forms normal and S-shaped distributions (Gould, 1969). Psychological barriers may affect Laggards (who latch onto information later) more than individuals of the Early Majority, and may influence the rate of diffusion spread. Additionally, one may have personal insecurities or concerns over privacy that may inhibit Twitter use. The decision to join or

participate in social media is a personal choice. Therefore, there might be some internal and psychological barriers that could prevent one from actively participating in Twitter.

The final barriers are physical, which are defined as blockades created by a location's landscape (Gould, 1969). Many landforms on Earth create physical, topographic, and naturally-forming barriers that can physically obstruct diffusion movement. Examples include lakes, oceans, hills, and mountains (Gould, 1969). If cellular or WIFI reception is not available because a location is remote, mountainous, or densely-forested, the physical landscape may lead to physical and technological barriers.

2.6 Quantitatively studying spatial diffusion

Researchers have used various approaches to quantify spatial diffusion. Hägerstrand used maps and diagrams to depict interactions between humans (Hägerstrand, 1967) and different social classes (Cliff, Pred, & Hägerstrand, 1992). Hägerstrand (1967) also took the quantitative approach of Monte Carlo experimentation in his famous work, *Innovation diffusion as a spatial process*. Monte Carlo experimentation is complex, and involves random sampling in different scenarios (Holmes, 2004). Hägerstrand's methods were especially impressive because all supporting mathematics were calculated without a computer (Cliff et al., 1992). Moreover, researchers including Rey, Murray, and Anselin (2011) and Gregory and Patuelli (2015) showed spatiotemporal diffusion using Moran's I statistics presented through scatterplots. These scatterplots were unique because locations were displayed along axes. Thus, these scatter plots, first conceptualized by Rey et al. (2011), can display diffusion processes over space and time. Poison distributions and Bayesian statistics were used by Acharya, Cao, Xu, Chen, and Pandit (2018) to study the diffusion of Dengue Fever. Lastly, some researchers have used the

Nearest Neighbor Index (NNI) and Standard Deviational Ellipses (SDEs) to show diffusion, and these are the methods I used in my thesis.

NNI can be used to quantitatively analyze the spatial diffusion of point data. Ecologists Philip Clark and Francis Evans developed and first published on NNI (Clark & Evans, 1954). NNI algorithms are based on two calculated distances. The algorithms first identify the distance between a data point and its closest adjacent point ("nearest neighbor"). I call this "Distance One." Then, the algorithms analyze a hypothetical random point data distribution. Next, calculations identify distances between a hypothetical point and its nearest neighbor. I call this "Distance Two." NNI is defined as the quotient of Distance One and Distance Two (Clark & Evans, 1954). Data are clustered and aggregated (dispersed and random) when NNI is closer to zero (one) (Clark & Evans, 1954). Data are uniform when NNI is closer to 2.1491 (Clark & Evans, 1954). Lee, Li, Wang, Wang, & Li (2020) applied NNI in their analysis of cancer (in Kaohsiung City, Taiwan in 2011), dengue fever (in Kaohsiung City, Taiwan from 2003 to 2008), crime (in Akron City, Ohio from 2008 to 2012), and 911 calls (in Portland, Oregon from 2013 to 2015). Wilkins and Ford (2007) showed data used in NNI studies are not limited to sociology, epidemiology, and human dynamics. Wilkins and Ford (2007) studied sand dunes formation and spread via NNI. Lee et al. (2014) used NNI and regression to categorize the spread of Targets, Walmarts, Dengue Fever cases, and mushroom farm locations. I explain their work in further detail in Chapter Three, as the methods used in Lee et al. (2014) are the foundation for the first part of my analysis.

Another quantitative approach to spatial diffusion is the SDE. Standard deviation, calculated mathematically as the square root of variance, indicates the proximity of data

around its mean (Sullivan & LaMorte, 2016). Low standard deviations (values generally less than one) suggest sample data are packed closer together, whereas higher values imply data points are farther apart (Sullivan & LaMorte, 2016). One, two, and three standard deviations are said to capture roughly 68%, 95%, and 99.7% of data, respectively (Galarnyk, 2018). Various GIS software can examine spatial point data by means of SDEs. This includes ArcGIS Pro, software offered through the Environmental Systems Research Institute (ESRI). “Directional Distribution (Standard Deviational Ellipse)” is the specific ArcGIS Pro tool used in my thesis (“How Directional Distribution (Standard Deviational Ellipse) works,” n.d.). Ellipses are drawn and scaled based upon standard deviations of x (latitude) and y (longitude) values within point data (“How Directional Distribution (Standard Deviational Ellipse) works,” n.d.).

CHAPTER THREE

DATA AND METHODS

3.1 Case studies

I researched two natural hazard case studies in the United States that varied in spatial size and temporal duration. The first case study was WSG, a winter storm that occurred 11 through 14 January 2019. It spanned 2.5k kilometers, and stretched from the Rockies to Mid-Atlantic (Jackson, 2019). WSG is considered a "synoptic" hazard, which includes meteorological events that range 100 to 5k kilometers in size. Temporally, synoptic events last up to a few weeks, but no less than a few days, in duration (DeCaria, n.d.). WSG formed from a Pacific Northwest low-pressure system on 9 January (Jackson, 2019). The system strengthened over the next 36 hours, and gained energy from a second low-pressure in the Gulf. Continental polar air from Canada, meeting the moisture from the Gulf, allowed for winter weather (Jackson, 2019). WSG reached maximum intensity on 13 January with hundreds of reported power outages, flight cancellations, and car accidents (Jackson, 2019). Central Missouri saw the highest snowfall totals, with reports nearing 51 centimeters (Jackson, 2019). I hypothesize WSG will demonstrate hierarchical and relocation diffusion, because weather information passes through various NOAA offices before reaching the general public (hierarchical diffusion). I also expect tweets to "follow" WSG's progression across the United States, which would suggest relocation diffusion.

My second case study is a tornado outbreak, in the Southeast, on 3 March 2019. The tornado outbreak was linked to a low-pressure system on the border of Texas and Louisiana. This low-pressure system was seen on the 1200Z (8:00 am Eastern) surface analysis map for 3

March (Santorelli, 2019b). Cold air from Missouri and Arkansas (Santorelli, 2019a) and warm air from Florida intersected over Alabama ("Tornadoes of March 3, 2019 Event Summary for Central Alabama," n.d.). Rotating winds and unstable air led to supercell storm formation ("March 3rd, 2019 Tornado Outbreak," 2019). Eighty tornadoes formed, according to preliminary data from the Storm Prediction Center (Storm Prediction Center, 2019). The outbreak caused 104 injuries ("Storm Prediction Center Storm Reports for 3/3/19," 2019). Twenty-three people died in one EF4 tornado that hit Beauregard and Smiths-Station (Lee County), Alabama ("Beauregard-Smiths Station Tornado - March 3, 2019," n.d.; Di Liberto, 2019). The NWS office in Birmingham, Alabama, reported this historic tornado was 1.6 kilometers wide, 43 kilometers long, and had maximum sustained winds of 274 kilometers per hour ("Beauregard-Smiths Station Tornado - March 3, 2019," n.d.). Its classification varies depending on the frame of reference. The entire outbreak area is of the synoptic scale because it is quite large and encompasses 1.7k kilometers, while an individual tornado is small and of the microscale. The tornado outbreak also developed and dissipated faster than WSG; thus, the two case studies have different spatiotemporal characteristics that lead to interesting analysis. I hypothesize the tornado outbreak will demonstrate contagious and expansion diffusion. People can discuss the tornado outbreak both in physical ("real world") and virtual (Twitter) space, which satisfies contagious diffusion's element of "direct contact." I predict more people will become aware of the tornado-outbreak during each day of the forecast period (28 February to 2 March), which demonstrates expansion diffusion.

3.2 Data

3.2.a Meteorological data

To accompany my analysis, I used various forms of meteorological data. I first gathered archived NWS watch, warning, and advisory shapefiles from the Iowa Environmental Mesonet (see Iowa State University, 2019). Watch, warning, and advisory shapefiles assisted in pre-processing WSG tweets, as described in greater detail in the "Filtering tweets using different approaches" section. I gathered two forms of Storm Prediction Center meteorological data for the tornado-outbreak case study: Convective Outlook maps and Preliminary Storm Reports. Convective Outlooks, which are issued daily, indicate locations of predicted severe convective weather such as thunderstorms and tornadoes. Convective Outlook categories range in order of least to most risk, as follows: "Thunderstorms," "Marginal," "Slight," "Enhanced," "Moderate," and "High." Generally, Convective Outlook maps display Thunderstorm as their largest risk area, and High risk indicates a particularly dangerous event for a smaller area. The second form of meteorological data for the tornado-outbreak case study, Storm Prediction Center Preliminary Storm Reports, are best estimates of tornado latitude and longitude ("Severe Weather Event Summaries," 2018). I use Storm Prediction Center Preliminary Storm Reports to compare tornado-outbreak tweets with reported tornado locations on 3 March 2019. While these data are only preliminary (2014 was the most recent year for available verified storm reports at the time of analysis), these preliminary values are suitable for the purposes of this thesis.

3.2.b Twitter data

I asked Dr. Yingjie Hu, Assistant Professor of Geography at the University of Buffalo, for assistance in collecting tweets. He used the Streaming API when we decided we would study

the event a couple of days in advance (the case for both studies), and the Searching API to add more days earlier in the forecast period prior to our decision to study the event (the case for WSG). He set a geographic boundary, reported in Figure 1, to constrain sampled tweet locations. I provided Dr. Hu with specific keywords to use in the APIs, listed in Table 1.

APIs returned 25,836 tweets related to WSG and 14,983 tweets related to the tornado-outbreak case study. I originally included action terms for the tornado tweets so that I could analyze shelter-seeking behaviors, but many tweets were noisy and off-topic. Some tweets mentioned "shelter," but were actually referring to animal shelters and pet adoptions. Additionally, tweets occasionally mentioned "weather" quite generally even though the phrase "severe weather" was searched in the APIs. For WSG, some tweets mentioned the word "winter," but were actually referring to an alcoholic beverage by the same name. Date ranges to retrieve tweets from the APIs also included additional days, both leading up to and following each case study. This proved it was necessary to filter tweets and remove irrelevant data as best as possible.

3.3 Filtering tweets using different approaches

3.3.a Filtering tweets by time and location

I used the Twitter Advanced Search (see "Advanced Search," 2020) to first gain a sense of when reputable sources tweeted about either case study. The Twitter Advanced Search can retrieve tweets based on keywords, date posted, username, and language ("Advanced Search," 2020). The Weather Channel has a massive presence in the weather community, and is now known for naming winter storms ("The Weather Channel Announces Winter Storm Names for

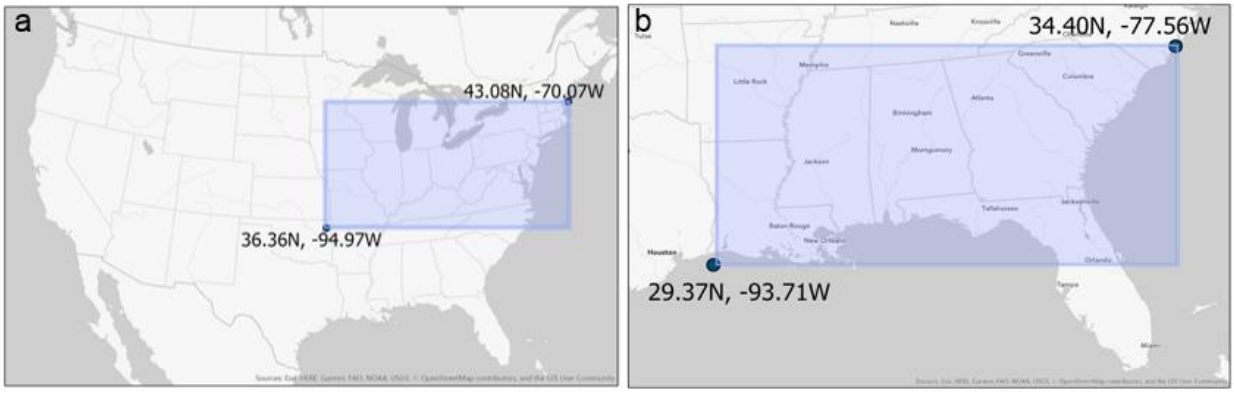


Figure 1: API sample regions. Geographic boundaries marking the latitude and longitude coordinates where the APIs searched for tweets in WSG (a) and the tornado outbreak (b).

Table 1: API keywords. Keywords entered for each case study into the API.

WSG Keywords	Tornado Outbreak Keywords
Winter Storm Gia	Tornado
Snow	Wind
Blizzard	Hail
	Funnel Cloud
	Severe weather
	Storm shelter
	Shelter
	Lowest level
	Basement
	Storm cellar
	Evacuate

for 2019-2020 Season," 2019). For these reasons, I searched The Weather Channel's Twitter account ("@weatherchannel") for the keyword "Winter Storm Gia." The Weather Channel hints to a developing system in this tweet from 7 January: *"A late-week snowmaker could impact parts South, Midwest and East…"* (@weatherchannel, 2019a). I compared the date of The Weather Channel's first tweet (7 January) to the date when NOAA's Weather Prediction Center ("@NWSWPC") first tweeted about a developing winter system. The Weather Prediction Center, another reliable source for weather information, first tweeted on 9 January: *"Moisture begins to increase by Friday [10 January 2019] ahead of a developing low pressure system in the Southern Plains..."* (@NWSWPC, 2019b). By 14 January, The Weather Channel began tweeting about a new winter storm: *"After the weekend's #WinterStormGia, another snowstorm is looking increasingly likely Friday [18 January 2019] -> next weekend in the Plains, Midwest and Northeast…"* (@weatherchannel, 2019b). The Weather Prediction Center last posted a tweet related to this case study on 13 January: *"#Snow is slowly tapering off over the Midwest and Ohio Valley but will continue into the evening over the interior Mid-Atlantic…"* (@NWSWPC, 2019a). I, therefore, used 7 to 14 January 2019 as the WSG date range for my thesis.

For the tornado outbreak case study, I searched Twitter's Advanced Search for the words "tornado," "tornadoes," "tornados," and "south east." I checked for tweets posted by Weather Underground ("@wunderground"), a sister company to The Weather Channel and another reliable source for weather information. Weather Underground first tweeted about potential severe weather on 28 February: *"Sunday [3 March 2019] brings another chance for #severe storms across the South as our Coast-to-Coast storm brings #snow to the Plains and Northeast…"* (@weatherunderground, 2019). The Weather Channel, which like Weather

Underground is also owned by IBM's The Weather Company, first reported the chance for severe weather in a tweet on 1 March: *"BREAKING: #WinterStormScott has been named and will bring heavy #snow from the Rockies into the Northeast with a #severe storm threat across parts of the South…"* (@weatherchannel, 2019c). The first tweet from the Storm Prediction Center was from 3 March, the day of the tornado outbreak itself: *"2:02pm CST #SPC Day 1 Outlook Enhanced Risk: across parts of southern alabama, the far western florida panhandle, southwest and south-central georgia…"* (@NWSSPC, 2019). The number of grammatical errors leads me to believe this was a tweet generated automatically. The Weather Channel indirectly mentioned the tornado-outbreak case study in an 8 March tweet: *"Our latest thinking on Saturday's [9 March 2019] #severe weather potential in the South, and whether Sunday's [10 March 2019] threat is anything to worry about for areas recovering from last Sunday's [3 March 2019] outbreak…"* (@weatherchannel, 2019d). However, I last found a tweet on 5 March that directly mentions the tornado outbreak: *"See the before-and-after images of the damage caused by Sunday's [3 March 2019] EF4 #tornado in Alabama...*" (@weatherchannel, 2019e). Therefore, 28 February through 5 March is the date range for the tornado-outbreak case study. The date ranges are summarized in Table 2.

I first filtered tweets by date posted. Originally, I had tweets from 5 to 30 January 2019 for the WSG case study, from 23 February to 15 March 2019 for the tornado-outbreak case study. I filtered based upon the dates listed in Table 2. Peak WSG intensity occurred from 11 to 14 January while 3 March is the date when 80 tornados formed; thus, date ranges from 7 to 14 January (WSG) and 28 February to 5 March (tornado outbreak) include days before, during, and

Table 2: Case study date ranges. Includes the forecast period, the event, and post-event tweets.

Case Study	Start Date (2019)	End Date (2019)
WSG	7 January	14 January
Tornado Outbreak	28 February	5 March

after both natural hazard events. Filtering by date reduced the tweet count from 25,836 to 5,053 (WSG) and from 14,983 to and 3,139 (tornado outbreak).

I then filtered tweets based on location. I only included tweets from within counties forecasted for either WSG or the tornado outbreak. For WSG, I first downloaded NWS watch, warning, and advisory shapefiles from Iowa State University's Iowa Environmental Mesonet. I then symbolized shapefiles (on ArcGIS Pro) so that it only included "winter weather advisory[ies]," "winter storm watch[es]," "winter storm warning[s]," and "blizzard warning[s]" from 8 to 15 January 2019. I used 8 January as the start date for identifying watches, warnings, and advisories because this corresponds to approximately 36 hours before WSG was at its peak. The NWS typically issues notices a few days prior to hazardous weather ("Watch/Warning/Advisory Definitions," n.d.). I used the ArcGIS Pro "Clip" tool to subset only tweets posted from 7 to 14 January and within locations warned for WSG by the NWS.

To filter tornado-outbreak tweets by location, I first downloaded Convective Outlook shapefiles from the Storm Prediction Center. I loaded Convective Outlook data into ArcGIS Pro for 3 March. I again used the "Clip" tool to locate tweets relevant to my case study. Tweets posted on 3 March within at least the Thunderstorm Convective Outlook risk category were included in this subsection. This process reduced the WSG sample size from 5,053 to 3,021 tweets and the tornado-outbreak sample size from 3,139 to 540 tweets.

3.3.b Filtering tweets by relevancy

I then manually filtered random samples of tweets for content. I used Python to randomly sample 500 WSG tweets, which was approximately 16% of WSG data. To be

consistent, I also randomly sampled 16% of the tornado-outbreak tweets which resulted in 90 tweets. I then separated the randomly sampled tweets into different categories. I first identified those tweets relevant to my thesis, and grouped these together in a category I called "Yes." I next indicated tweets possibly relevant to my thesis, and called this category "Maybe." I grouped tweets irrelevant to my thesis in a "No" category. Tweets classified in the "Maybe" list needed additional consideration. Upon reexamination, "Maybe" tweets were either moved to "Yes" or "No" lists in an attempt to accurately classify tweets.

I created Python graphs to display the 20 most common words in the "Yes," "Maybe," and "No" lists to identify words most commonly used in irrelevant tweets. I removed punctuation, URLs, duplicate tweets, and stop words prior to graphing word frequencies. Stop word lists included API keywords used to acquire tweets (such as "Winter Storm Gia" and "tornado"). All characters in randomly sampled tweets were tokenized and converted to the lower case, which is a standard text-cleaning process. Figure 2 displays the 20 most common words in the WSG "No" list. I then removed any WSG tweet (whether in the 500 tweet random sample or not) that contained at least one of the following words: "Skiing," "skiing," "Drinking," "drinking," "Mountain," "mountain," "Ski," "ski," "Tubing," "tubing," "Snowboarding," "snowboarding," "Hops," "hops," "Snowboard," and "snowboard." Notice how words such as "day," "today," "first," "park," "12," "river," "last," "like," "Colorado," "us," "5," and "35" appeared in Figure 2, but were not included in the list to remove tweets. I considered tweets containing any of these additional words as possibly related to WSG. Filtering in this way reduced the WSG tweet sample size from 3,021 to 2,881. Finally, I removed any duplicate tweets, which further reduced the WSG sample size to 2,844 tweets.

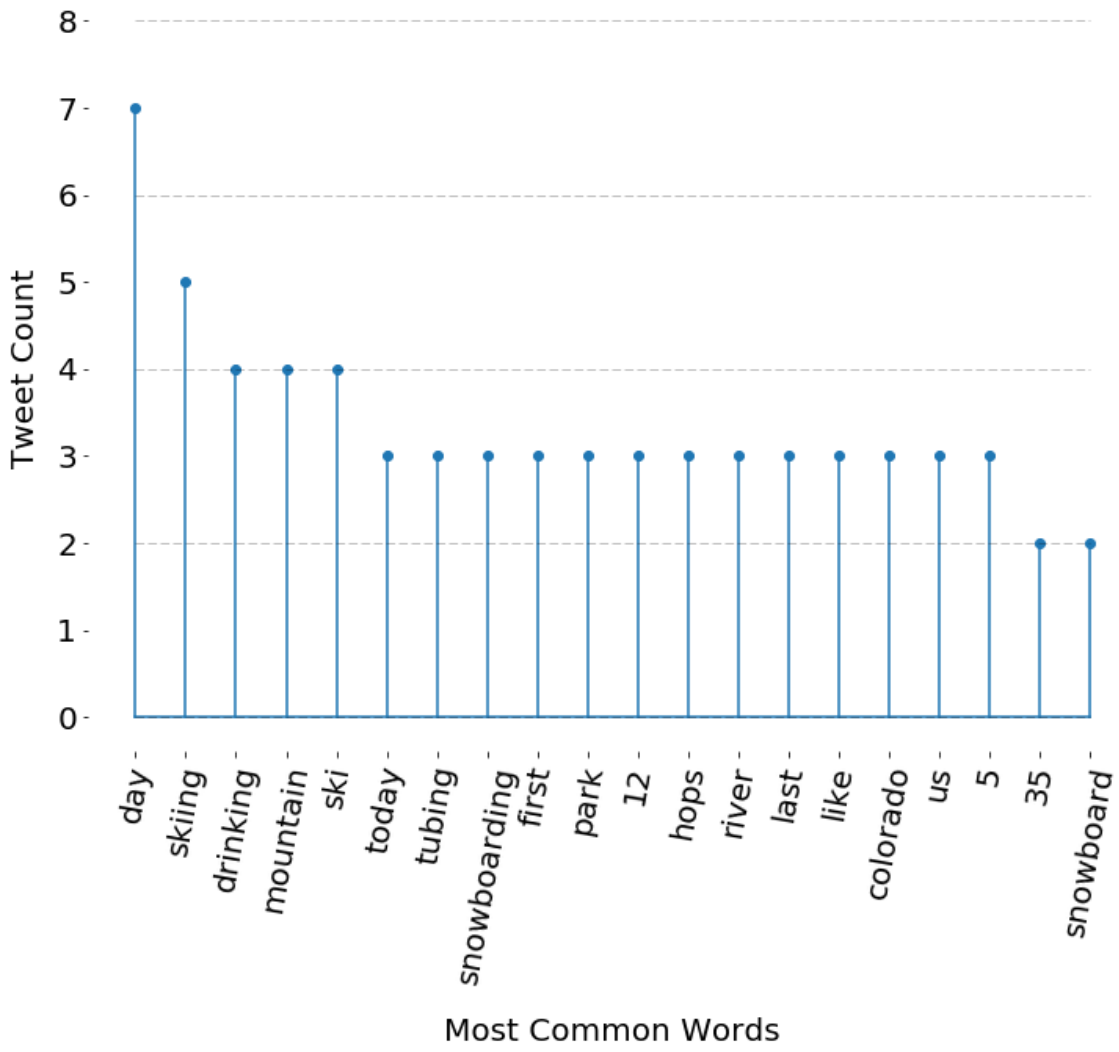


Figure 2: Irrelevant WSG tweets. The 20 most common words from irrelevant WSG tweets within the sample of 500 random tweets.

Using a similar approach, I next created Python graphs to display the 20 most common words in the "Yes," "Maybe," and "No" tornado-outbreak tweets. I removed punctuation, URLs, duplicate tweets, and stop words (with the addition of API keywords used to acquire tweets such as "tornado" and "shelter"). I tokenized and converted all tweets to lowercase and created my graphs. I removed tornado-outbreak tweets that contained certain words that are identified in Figure 3. I did not include the words "pressure," "feels," "gust," "weather," "nne," "rain," "temperature," "n," "today," "104mph," "temp," "89mph," "360f," "000in," "93," "light," "11mph," "86," and "361f" in my criteria of tornado-outbreak tweets to remove. It seems many of these phrases were the result of automated weather station readings that happen automatically report to Twitter. I found these tweets problematic because auto-generated tweets are not from actual people, and could possibly create unreliable data. Automated weather stations almost always report humidity. I used this to my advantage, and removed any tornado-outbreak tweet (whether in the 90 tweet sample or not) that contained the word "humidity." The filtering process reduced the tornado-outbreak tweet sample size from 540 to 330.

Skimming the remaining tornado tweets showed they still contained irrelevant data; thus, I manually filtered the tornado outbreak-tweets a second time. I did not complete a second manual filter on the WSG tweets. I randomly sampled 90 tweets from the 330 tornado outbreak tweet sample, and classified the tweets as "Yes," "Maybe," and "No." Figure 4 shows the updated 20 most common words in the "No" list. I additionally removed tornado-outbreak tweets (whether in the 90 tweet sample or not) that contained the words "Pressure,"

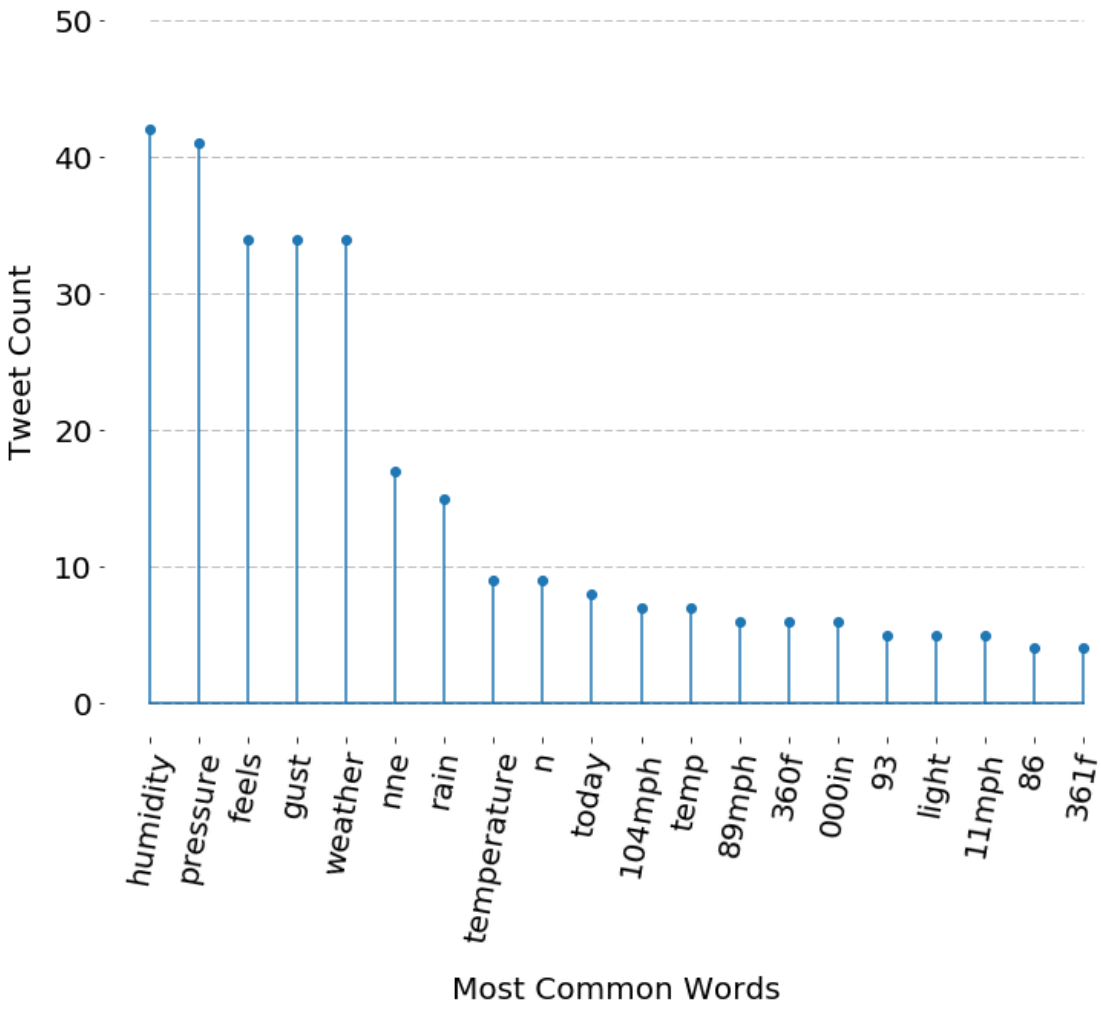


Figure 3: Irrelevant tornado-outbreak tweets (first filter pass). The 20 most common words from irrelevant tornado-outbreak tweets within the sample of 90 random tweets.

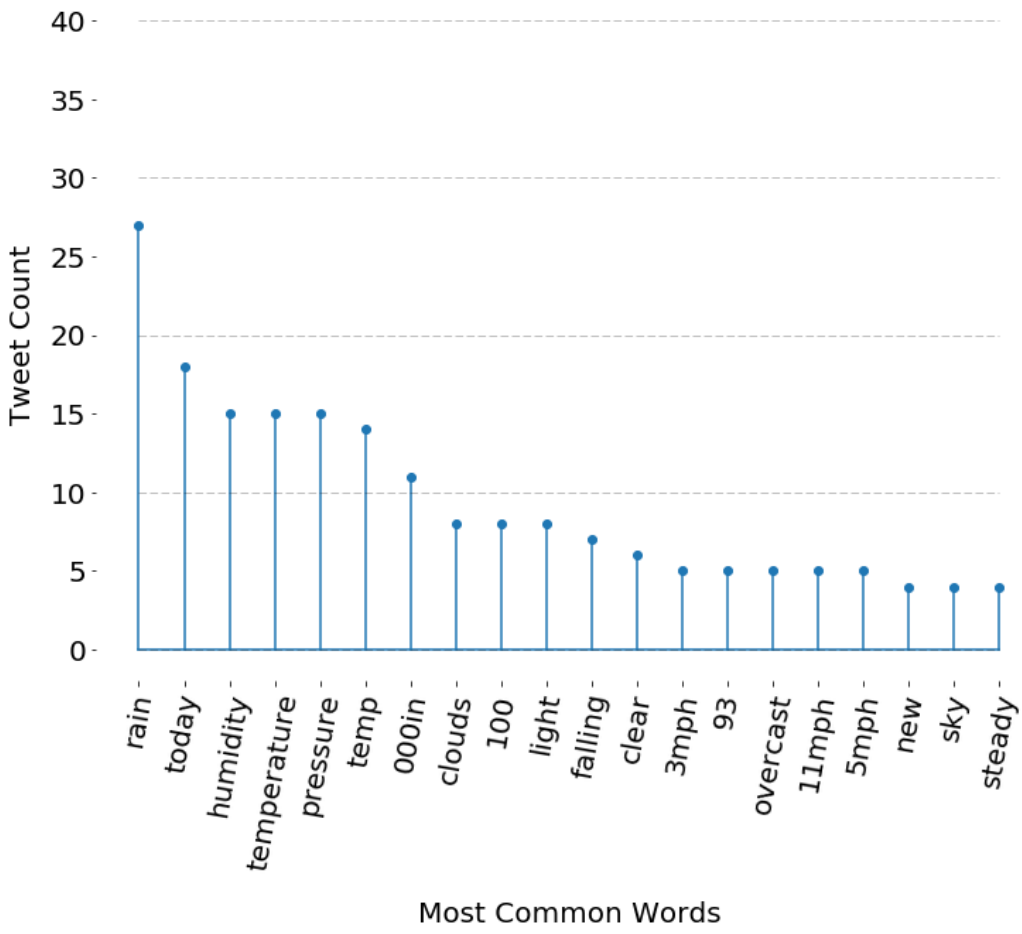


Figure 4: Irrelevant tornado-outbreak tweets (second filter pass). The 20 most common words from 90 irrelevant tornado outbreak tweets within the 330 tornado outbreak tweet random sample.

"pressure," "Temp," and "temp." As a result, the tornado-outbreak sample size was reduced from 330 to 217 tweets. Removing duplicate tweets further reduced the sample size to 213.

3.4 Nearest Neighbor Index (NNI)

One technique I used to analyze data was the Nearest Neighbor Index (NNI), which I used as the response variable in the regression model, described in detail in the following methods subsection. NNI relies on the statistical concept "nearest neighbor." Nearest neighbor calculates the distance between any arbitrary point and its closest "neighboring" point. The overall NNI calculation uses the average calculated nearest neighbor distance, the numerator of Equation One.

$$NNI = \frac{\sum \frac{Neighboring\ distance\ between\ points\ in\ data}{Number\ of\ points\ in\ data}}{\frac{\sqrt{\frac{Hypothetical\ data's\ area}{Number\ of\ points\ in\ hypothetical\ data}}}{2}}$$

Equation One

NNI is therefore a ratio and a unitless value. It compares the average distance between points to the average distance of arbitrary points.

I used a "for" loop in R, alongside the "nni" function (Evans, n.d.) to calculate NNI values for my Twitter data. I first stored tweet latitude and longitude values in a spatial data frame. I took the first tweet's coordinates and calculated NNI. Since the system contained only one point, NNI resulted in an "infinite" value. Next, I added a second tweet and calculated NNI. I then calculated NNI after adding a third tweet, a fourth tweet, a fifth tweet, and so on through all tweets in the data frame. I used the following line of code to produce these NNI values:

$$for\big(i\ in\ 1{:}\ length(data_frame)\big)\{list\big[[i]\big] = nni(data[1{:}\ i,\ ], win = extent)\}$$

I followed these steps in both case studies, and saved calculated NNI values in a data frame for analysis.

3.5 NNI and regression analysis

I used NNI values in a regression model based on methods used by Lee et al. (2014), who evaluated the effectiveness of spatiotemporal diffusion regression models for point data. For their work, Lee et al. (2014) examined the spread of mushroom farms, Dengue Fever, Targets, and Walmarts over space and time. NNI was the response (dependent) variable while time was the exploratory (independent) variable. They modeled the data using different base equations such as linear, quadratic, cubic, inverse, power, logarithmic, logistic, S-curve, compound, exponential, and growth (see Lee et al. (2014) chart on page 6). Next, they compared coefficient of determination (R^2) values across regression model runs to determine what types of spatiotemporal diffusion classifications (contagious, expansion, hierarchical, and relocation) were best represented by each equation. Lee et al. (2014) attributed higher R^2 values as increased assurance that any arbitrary spatiotemporal case study can be modeled by a respective regression base equation. Contagious and expansion diffusion events were found to be best modeled via an inverse base regression equation of the form:

$$Y = b_0 + \frac{b_1}{t}$$

Conversely, hierarchical and relocation diffusion events were found to be best modeled by a cubic base regression equation:

$$Y = b_0 + b_1 t + b_2 t^2 + b_3 t^3$$

I used similar methods to determine how spatiotemporal tweet diffusion varies between WSG and the tornado-outbreak case study. Unlike the examples presented by Lee et al. (2014), I did not first associate case studies with diffusion classifications (contagious, expansion, hierarchical, and relocation) prior to data analysis; instead, I aim to use the findings by Lee et al. (2014) to assign a diffusion classification suggested by the regression results.

For each of the case studies, I cleaned the NNI data using R's Tidyverse. Next, I calculated elapsed time from any arbitrary tweet and the first tweet. I then converted elapsed time from seconds to hours to more easily interpret results. I removed "infinite" and "not a number" (NaN) values. Clark and Evans (1954) report NNI values less (more) than one indicate clustered (dispersed) data, while NNI values close to 2.1491 indicate uniform data. I removed outliers, by means of two standard deviations, prior to performing regression. There were 35 outliers in the WSG case study, and 6 for the tornado outbreak. I then used the following line of R code to run cubic regression models:

$$lm(NNI \sim poly(hours, 3, raw = TRUE))$$

The "poly" function is used for polynomial regression, in this case, cubic polynomial regression (as indicated by the number three). I made "NNI" and "hours" the response and exploratory variables, respectively. The parameter "raw = TRUE" prevents calculation using orthogonal polygons (R-core R-core@R-project.org, n.d.). I ran inverse regression models using this code line:

$$lm(NNI \sim I(1/hours))$$

The “Inhibit Interpretation/ Conversion of Objects” (“I”) function (“Inhibit Interpretation/Conversion of Objects,” n.d.) prevents binary division, which is what the “/” symbol means to R (“R Language Definition,” 2020). Binary division would be problematic because values in “NNI” and “hours” are numeric (base ten) not binary (base two). The “augment” function in the “broom” package (Hayes, n.d.) reports regression model standard error, fitted values, and residuals among other parameters for assessing model quality:

$$broom :: augment(regression_model)$$

3.6 Standard Deviational Ellipse (SDE)

I used SDE as an alternative method to quantitatively study spatiotemporal diffusion. I used the “Directional Distribution (Standard Deviational Ellipse)” tool in ArcGIS Pro to map SDEs atop case study tweet layers. This tool provided detailed SDE properties such as “Mean Center Location,” “Standard Distance,” “Rotation,” “Shape Length,” and “Shape Area.” “Mean Center Location” is defined by X and Y coordinate centroids. Standard deviation, along the X and Y axes, is how to interpret “Standard Distance.” Standard distance is particularly useful because it defines the SDE’s overall shape and is not effected by the SDE’s area. “Rotation” defines the SDE’s positioning and alignment, whether titled horizontal or more vertical. “Shape Length” and “Shape Area” report SDE length and area, respectively.

I used “Mean Center Location” and “Shape Area” in my analysis for each case study. I then used R to calculate how the SDE’s X component, Y component, and Shape Area changed throughout case studies.

$$for(i\ in\ 1{:}\ Number_of_points_in_data_frame)\ \{list[i] = data[i+1] - data[i]\}$$

The phrase "$data[i+1]$" was used to access data (whether X component, Y component, or Shape Area) stored one time step in the future while the phrase "$-\ data[i]$" subtracted this value from data stored in the current time step. The difference shows the change in SDE components, and provides information about how the tweets diffused over time.

CHAPTER FOUR

RESULTS

4.1 Daily tweet counts over case study periods

I noticed similar, yet uniquely distinguishable, temporal tweet patterns across both case studies. Figure 5a shows daily WSG tweet counts from 7 January to 14 January 2019. Tweets remained fairly steady across the first two days in the WSG case study, as there was only a difference of three tweets between 7 January and 8 January (69 and 71 tweets on each day, respectively). The Weather Channel had only just begun to report a potential winter storm. Tweet counts reached 172 on 9 January, but dipped to 93 on 10 January. Counts then rose quite dramatically—by 850%—from 10 January (93 tweets) to 12 January (884 tweets). WSG traversed through the Central Plains and Ohio River Valley on these days, and caused numerous flight cancelations (Jackson, 2019). Tweet counts again rose between 12 and 13 January, but at a less steep slope than previously observed. These days, 12 and 13 January, saw the highest counts seen across any two-day period. Figure 6 shows a comparison of tweet locations during the forecast period before the event (6a) versus midway through WSG (6b). Tweet counts plummeted from 934 on 13 January to 296 on 14 January. By this point in the storm's track, WSG had already passed peek intensity. The system was dissipating, and social media interest had already moved onto the next winter storm, as shown by a tweet saying, *"After the weekend's #WinterStormGia another snowstorm is looking increasingly likely Friday [18 January 2019] -> next weekend in the Plains, Midwest and Northeast…"* (@weatherchannel, 2019b).

The daily tweet frequency during the tornado-outbreak case study exhibited a different pattern (Figure 5b). Only 15 tweets were observed on 28 February—the lowest single-day total

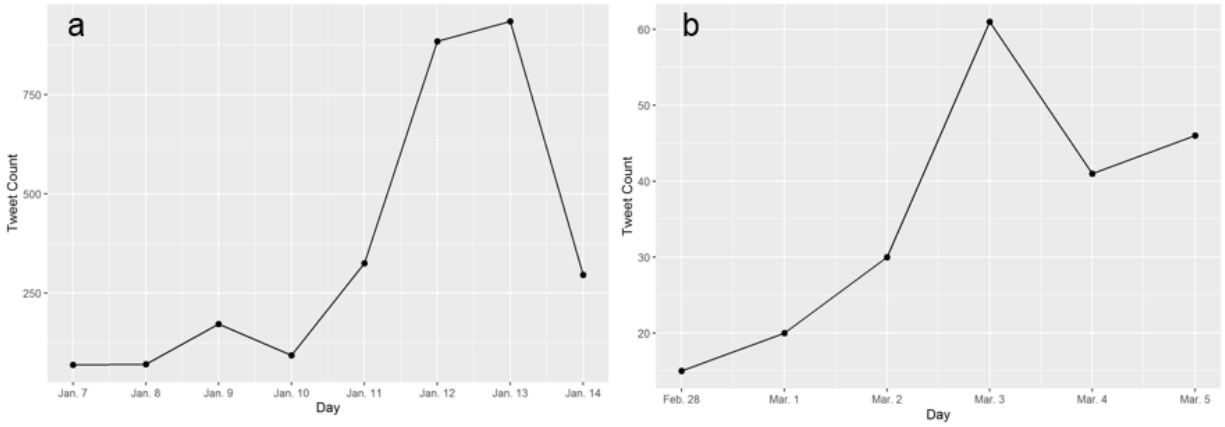


Figure 5: Daily tweet counts. Daily WSG (a) and tornado-outbreak (b) tweet counts.



Figure 6: WSG tweets on two different days. WSG tweets, shown by blue dots, for 7 January 2019 (a) and 12 January (b). WSG snowfall was most widespread from 11 to 14 January 2019. There are noticeably more tweets on 12 January than 7 January 2019.

seen across my entire thesis. There was no pre-event dip in tweet counts for the tornado-outbreak case study, as observed in WSG between 9 and 10 January. Tweet counts steadily rose between 1 March (20 tweets) and 2 March (30 tweets), as word of a possible tornado outbreak spread through news sources such as The Weather Channel. Figure 7 compares the location of tweets on 28 February (7a), when Weather Underground first mentioned the tornado outbreak, versus a few days later on 1 March (7b). Tweet counts reached a maxima on 3 March (60 tweets). The Storm Prediction Center reported over 80 tornadoes on this day. Tweet counts dropped considerably from 3 March (60 tweets) to 4 March (41 tweets), but rose again slightly on 5 March (46 tweets). WSG counts, conversely, never increased post-event and instead dropped dramatically.

4.2 Spatiotemporal diffusion regression analysis

For my regression analysis, NNI and time are the response (dependent) and explanatory (independent) variables, respectively. Figure 8 shows NNI plotted against elapsed time (in hours) since the first WSG tweet (Δt). The purple line in Figure 8a shows a cubic regression fit for WSG, which is modeled by Equation Two and resulted in an R^2 value of 0.836.

$$NNI = (4.85e-01) - (5.13e-03)\Delta t + (3.88e-05)\Delta t^2 - (1.11e-07)\Delta t^3 + \varepsilon$$

Equation Two

Similar to the approach of Lee et al. (2014), I compared R^2 values across different regression model runs. The WSG case study was not well-modeled by inverse regression (Figure 8b), as the R^2 value was lower at 0.459. Equation Three is the "line of best fit" for the WSG inverse regression model.

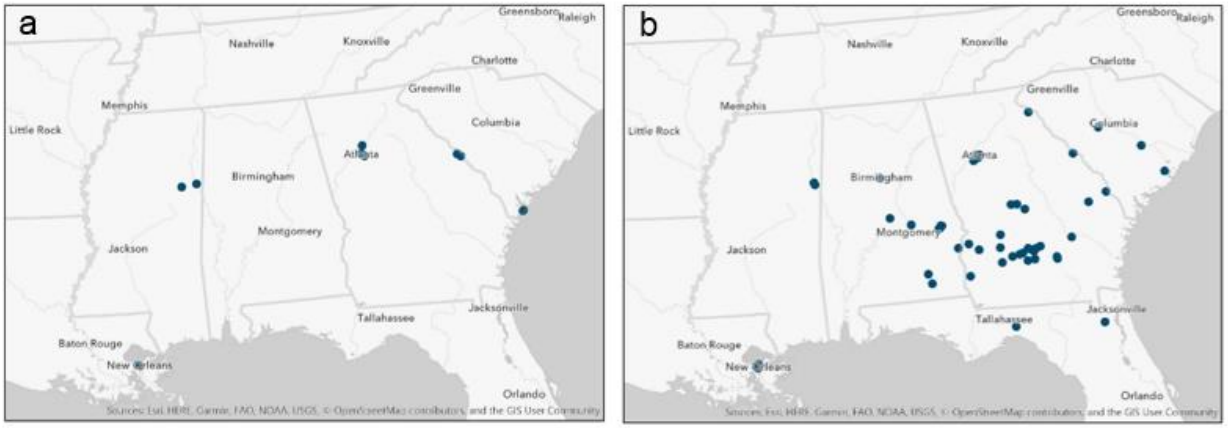


Figure 7: Tornado-outbreak tweets on two different days. Tornado-outbreak tweets, showed by blue dots, on 28 February 2019 (a) versus 1 March 2019 (b). Atmospheric conditions leading to the tornado outbreak started to form on 1 and 2 March 2019. There are noticeably more tweets on 1 March than 28 February 2019.

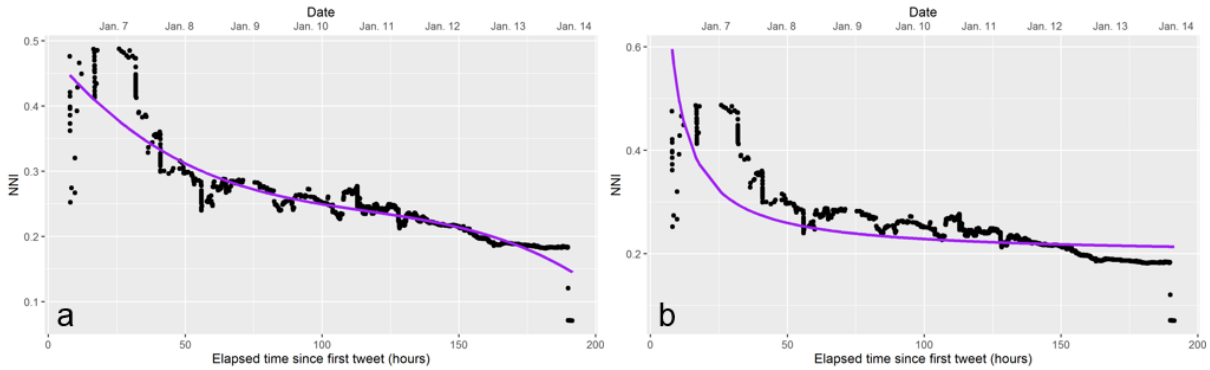


Figure 8: WSG regression analysis. NNI of WSG tweets (points) versus elapsed time. Purple line models cubic regression (a) and inverse regression (b).

$$NNI = (1.97e - 01) + (3.13)\frac{1}{\Delta t} + \varepsilon$$

Equation Three

Cubic and inverse regression were not suitable models for the tornado-outbreak case study. The cubic model, for instance, had a R^2 value of 0.159. The purple line, seen in Figure 9a, poorly fits the data. This lack of fit is especially evident between hours 0 and 25, along the x axis. Data are widely scattered both above and below the purple "line of best fit," which is modeled by the following equation:

$$NNI = (5.13e - 01) - (2.90e - 04)\Delta t - (2.47e - 05)\Delta t^2 + (1.49e - 07)\Delta t^3 + \varepsilon$$

Equation Four

The inverse regression model performed worse (R^2 value 0.123) than the cubic model, as seen in Figure 9b. Equation five produces the "line of best fit" for the inverse regression model:

$$NNI = (3.89e - 01) + (1.62)\frac{1}{\Delta t} + \varepsilon$$

Equation Five

4.3 Standard Deviational Ellipse (SDE) analysis

I used the "Directional Distribution (Standard Deviational Ellipse)" tool in ArcGIS Pro to analyze changes in SDE location and area over two-day periods. Results are best conceptualized with respect to the Cartesian coordinate system. The x-axis can be broken up into "positive-x" and "negative-x" while the y-axis can be broken into "positive-y" and "negative-y" axes (Figure 10).

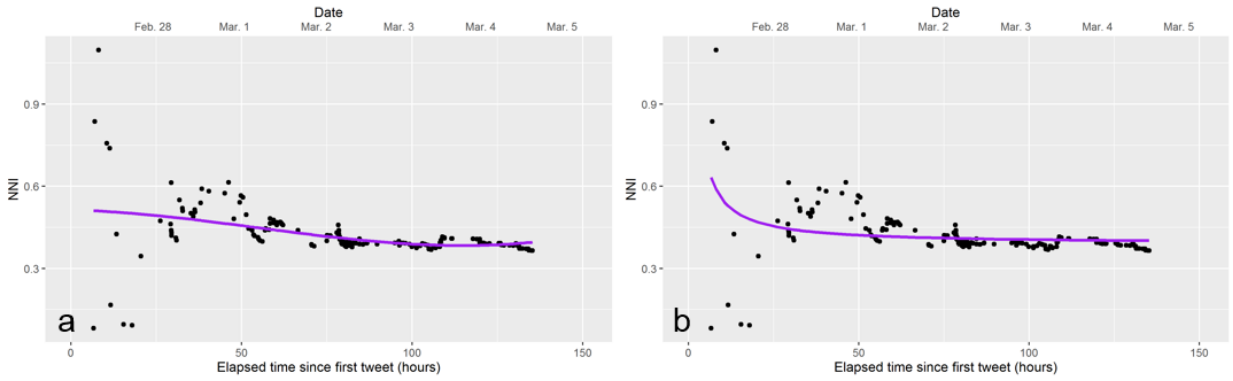


Figure 9: Tornado-outbreak regression analysis. NNI of tornado-outbreak tweets (points) versus elapsed time. Purple line models cubic regression (a) and inverse regression (b).

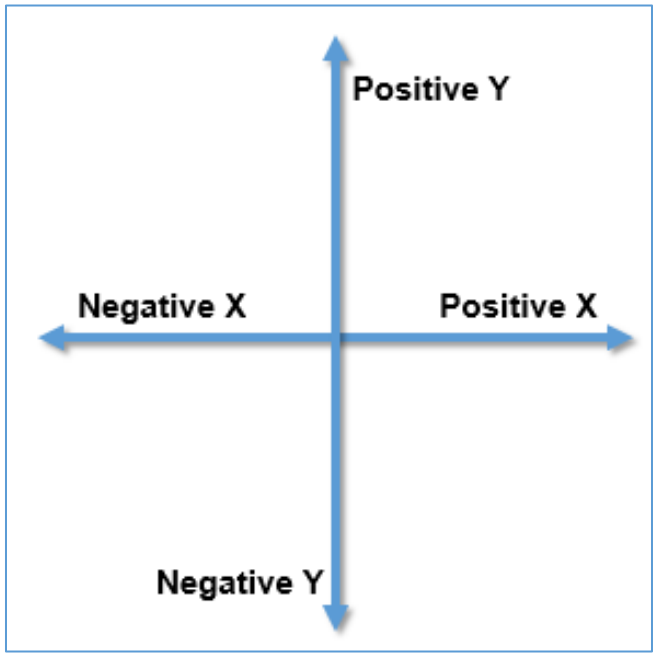


Figure 10: SDE mean center location reference axes.

Table 3 reports changes in SDE mean center locations for the WSG case study. Locations shift in both positive and negative directions along the X and Y axes, described in Figure 10. Between 7 and 8 January (first row in Table 3), SDE mean center location shifted 150 km in the positive-x direction and 19 km in the negative-y direction. Then, between 8 and 9 January (second row in Table 3), SDE mean center location shifted 174 km in the positive-x direction and 45 km in the positive-y direction. The remaining rows in Table 3 can be interpreted similarly. The first row (7 to 8 January) is the only instance when one X or Y column increases while the other decreases. Across other time periods, one component (either X or Y) increased while the other component (either X or Y) decreased. Mean center locations, throughout WSG, are mapped in Figure 11. Table 4 shows the change in SDE area for the WSG case study. The SDE decreased in size and became more concentrated from 7 to 8 January (change in SDE shape area of -3.8+E08) and from 8 to 9 January (change in SDE shape area of -2.7E+08). SDE area increased over the next few days before decreasing again from 11 to 12 January. Figure 12 shows where tweets were located each day, along with SDEs (purple polygons) and mean centers (purple dots).

Similar to WSG, I observed both positive and negative changes in mean center location and SDE area for the tornado-outbreak case study. Some days had positive changes in both X and Y axes while other days showed negative X and Y axes changes, as seen in Table 5. The date range of 3 to 4 March was the only time when one component (either X or Y mean center location) increased while the other decreased. I observed this general trend, of values both increasing and decreasing throughout time, when analyzing SDE area too. SDE mean center locations throughout the tornado-outbreak case study are mapped in Figure 13. SDE area

Table 3: Properties of SDE mean centers for WSG tweets. “X (km)” and “Y (km)” are the X and Y components to the location of the mean center. “X Standard Distance (km)” and “Y Standard Distance (km)” are standard deviations along the X and Y axes. “Delta X (km)” and “Delta Y (km)” are the change in X and Y components.

Date Range	X (km)	Y (km)	X Standard Distance (km)	Y Standard Distance (km)	Delta X (km)	Delta Y (km)
7 Jan	1,054	364	1487	206	150	-19
8 Jan	1,203	345	933	197		
8 Jan	1,203	345	933	197	174	46
9 Jan	1,377	391	554	179		
9 Jan	1,377	391	554	179	-201	-61
10 Jan	1,176	329	817	218		
10 Jan	1,176	329	817	218	-412	-62
11 Jan	764	267	1040	234		
11 Jan	764	267	1040	234	145	30
12 Jan	909	298	918	192		
12 Jan	909	298	918	192	296	20
13 Jan	1,206	318	982	166		
13 Jan	1,206	318	982	166	-44	-4
14 Jan	1,162	314	1215	429		

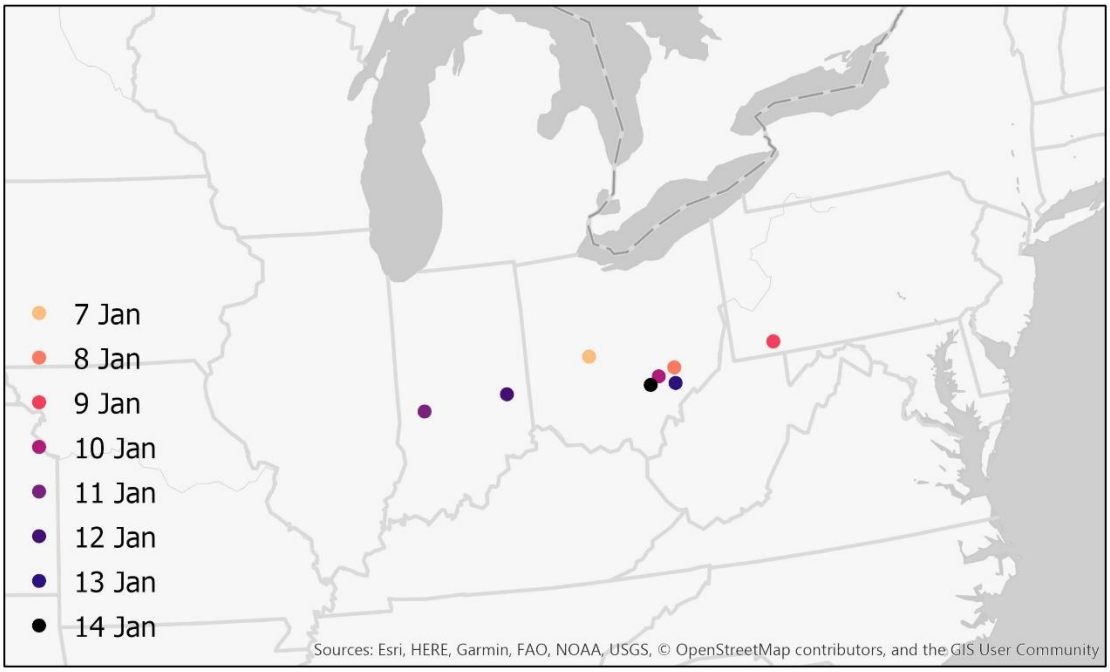


Figure 11: Daily mean center locations for WSG tweets.

Table 4: Properties of SDE area for WSG tweets.

Date Range	Shape Area (km)	X Standard Distance (km)	Y Standard Distance (km)	Change in Shape Area (km)
7 Jan	9.6E+08	1487	206	-3.8E+08
8 Jan	5.8E+08	933	197	
8 Jan	5.8E+08	933	197	-2.7E+08
9 Jan	3.1E+08	554	179	
9 Jan	3.1E+08	554	179	2.5E+08
10 Jan	5.6E+08	817	218	
10 Jan	5.6E+08	817	218	2.1E+08
11 Jan	7.6E+08	1040	234	
11 Jan	7.6E+08	1040	234	-2.1E+08
12 Jan	5.5E+08	918	192	
12 Jan	5.5E+08	918	192	-4.0E+07
13 Jan	5.1E+08	982	166	
13 Jan	5.1E+08	982	166	1.1E+09
14 Jan	1.6E+09	1215	429	

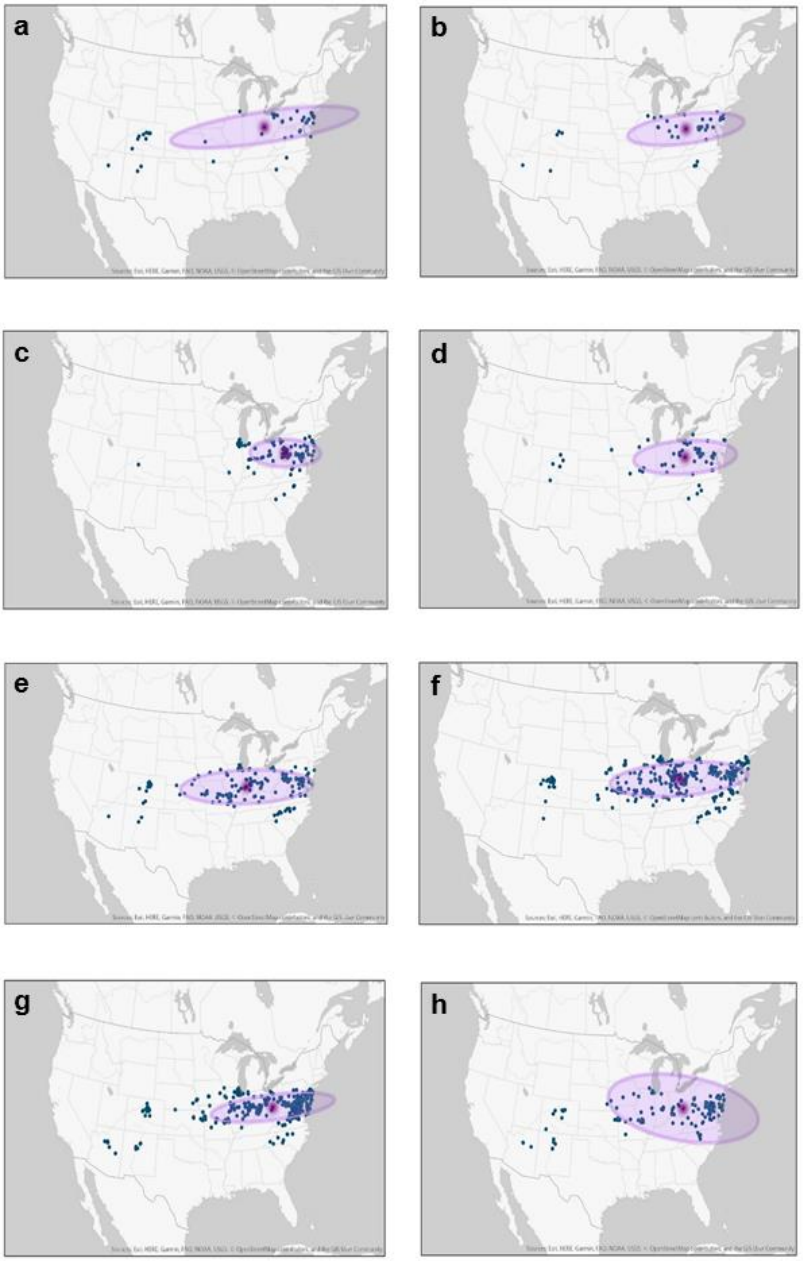


Figure 12: WSG SDE maps. WSG SDE map for 7 January (a), 8 January (b), 9 January (c), 10 January (d), 11 January (e), 12 January (f), 13 January (g), 14 January (h) 2019 where blue points are tweets, purple polygons are SDEs, and purple points are mean center locations.

Table 5: Properties of the SDE mean centers for the tornado-outbreak tweets. “X (km)” and “Y (km)” are the X and Y components to the location of the mean center. “X Standard Distance (km)” and “Y Standard Distance (km)” are the standard deviations along the X and Y axes. “Delta X (km)” and “Delta Y (km)” are the change in X and Y components.

Date	X (km)	Y (km)	X Standard Distance (km)	Y Standard Distance (km)	Delta X (km)	Delta (km)
28-Feb	1,041	-443	423	167	-105	-130
1-Mar	936	-573	553	283		
1-Mar	936,472	-573	553	283	191	162
2-Mar	1,128	-410	584	176		
2-Mar	1,128	-410	584	176	-30	-114
3-Mar	1,098	-524	316	159		
3-Mar	1,098	-524	316	159	-119	22
4-Mar	980	-502	488	161		
4-Mar	980	-502	488	161	49	53
5-Mar	1,028	-449	405	161		

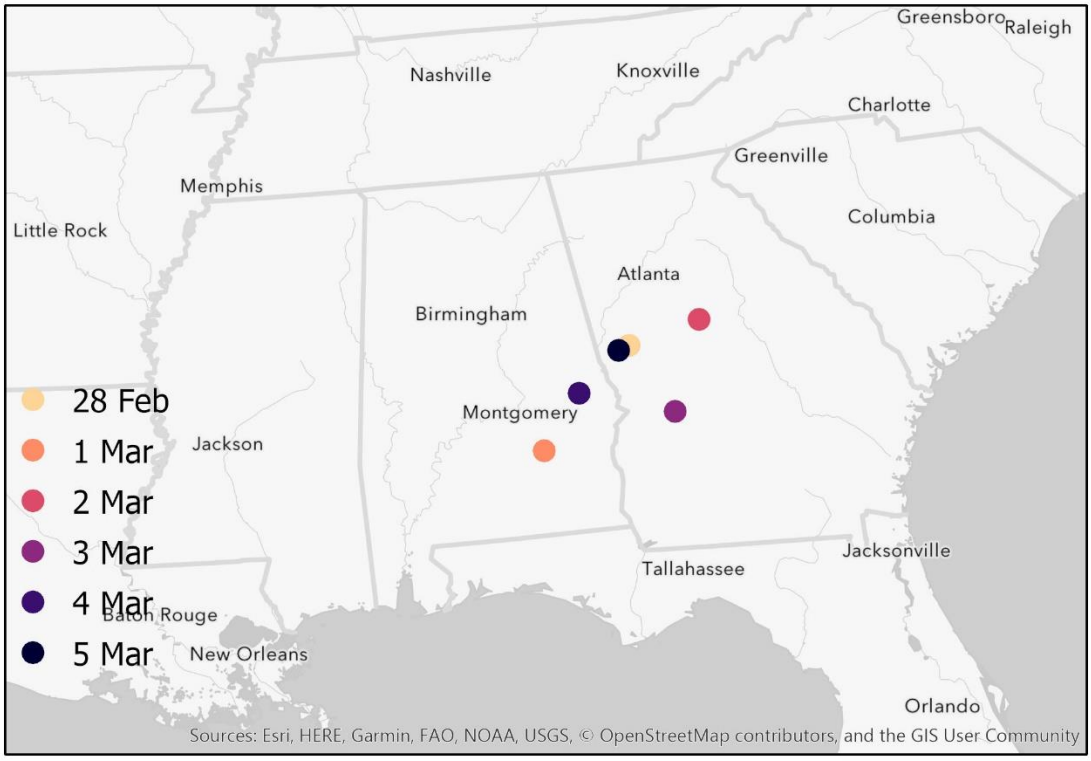


Figure 13: Daily mean center locations for tornado-outbreak tweets.

increased over some date ranges and decreased over others. The increase in SDE area on 3 to 4 March is the most notable. The SDE is smaller and more concentrated over southeastern Alabama and central Georgia on 3 March. By 4 March, the SDE is more expansive, spreading from Mississippi to central South Carolina, as seen in Table 6. Figure 14 shows where tweets were located each day, along with SDEs (purple polygons), and mean centers (purple dots).

Table 6: Properties of SDE area for the tornado-outbreak tweets.

Date	Shape Area (km)	X Standard Distance (km)	Y Standard Distance (km)	Change in Shape Area (km)
28-Feb	2.2E+08	423	167	2.71E+08
1-Mar	4.9E+08	553	283	
1-Mar	4.9E+08	553	283	-1.69E+08
2-Mar	3.2E+08	584	176	
2-Mar	3.2E+08	584	176	-1.65E+08
3-Mar	1.6E+08	316	159	
3-Mar	1.6E+08	316	159	8.80E+07
4-Mar	2.5E+08	488	161	
4-Mar	2.5E+08	488	161	-4.10E+07
5-Mar	2.1E+08	405	161	

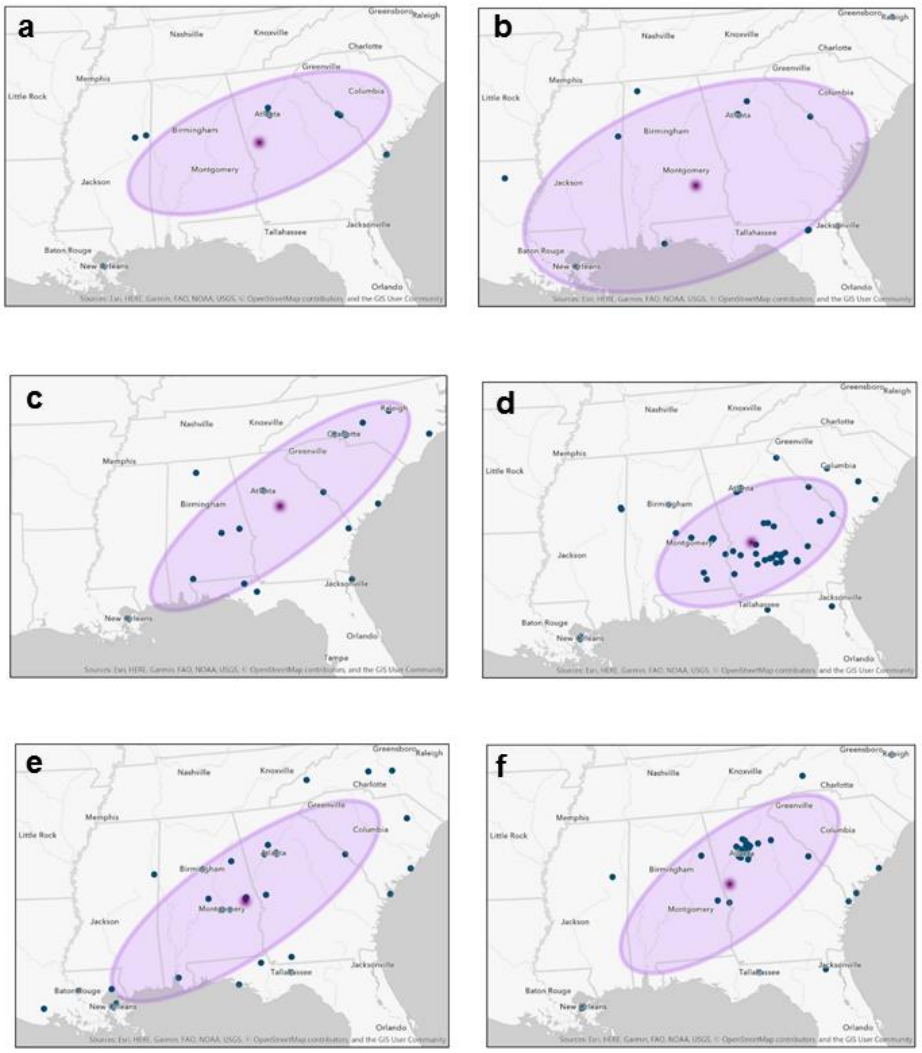


Figure 14: Tornado-outbreak SDE maps. Tornado-outbreak SDE maps for 28 February (a), 1 March (b), 2 March (c), 3 March (d), 4 March (e), and 5 March 2019 (f) where blue dots are tweets, purple polygons are SDEs, and purple points are mean center locations.

CHAPTER FIVE

DISCUSSION

5.1 WSG diffusion over space and time

WSG daily tweet counts mirrored what Kryvasheyeu et al. (2016) observed in their analysis of Superstorm Sandy. Superstorm Sandy tweet counts increased as the system strengthened and news media began reporting its potential impacts (Kryvasheyeu et al., 2016). These Twitter users are the Early Adopters and Early Majority, described by (Gladwell, 2002). Tweet counts peaked as Sandy made landfall, and decreased shortly after (Kryvasheyeu et al., 2016). Similarly, WSG tweet counts generally increased in days leading up to the event (7 to 10 January 2019); however, there was one anomaly from 8 to 9 January when tweet counts dropped. I speculate two reasons for the anomaly: one, people possibly dismissed warnings about a developing winter storm; or on the complete opposite spectrum, people were preparing for WSG and did not have time to tweet. A thorough tweet content analysis would be needed to determine factors leading to the dip, while a statistical analysis would be needed to verify significance. WSG tweets peaked on 13 January, during the storm's most intense period from 11 to 13 January 2019. Perhaps tweet counts were high due to the system's vast spatial extent from the Rockies to the Mid-Atlantic, which would cause many people to simultaneously live through WSG, and this shared experience perhaps encouraged people to tweet. People seemed to quickly move onto other pursuits after WSG, as tweet counts dropped steeply from 13 to 14 January. The Weather Channel immediately began tweeting about the next winter storm, *"After the weekend's #WinterStormGia another snowstorm is looking increasingly likely*

Friday [18 January 2019] -> next weekend in the Plains, Midwest and Northeast…"

(@weatherchannel, 2019b).

Locations of WSG tweets did not follow the storm's progression. WSG formed over the Mid-West on 11 and 12 January. The system reached the Ohio River Valley on 13 January, and Mid-Atlantic hours later. WSG moved offshore starting 14 January, and completely dissipated by 15 January (Jackson, 2019). I expected tweet concentrations to follow the storm's path, from the Rockies to the Mid-Atlantic, and thereby demonstrate relocation diffusion. However, regardless of the day, I observed high concentrations of tweets over the Ohio River Valley and Mid-Atlantic. These regions have a mix of urban and rural settings, but high population densities in cities such as St. Louis, Indianapolis, Pittsburgh, and Washington D.C. It is likely these cities influenced mean center locations, presented in Figure 12. Individuals who make up Twitter's young (18 to 24) and educated (college-level) demographic (Perrin & Anderson, 2019) often live and congregate within cities. Population dynamics likely pulled SDE mean center locations towards the heavily populated east coast, and also affected the diffusion of tweets because they were always centered in populated areas.

5.2 Tornado-outbreak tweets over space and time

Tweet counts from the tornado-outbreak case study also followed observations of Sandy described by Kryvasheyeu et al. (2016), as tweet counts increased in days leading up to the event (28 February to 2 March 2019) and decreased immediately following peak tweets on 3 March 2019. However, different from WSG, tweet counts slightly rebounded afterwards on 5 March. I observed the general pattern in tweet counts rather than examining the rebound's statistical significance. Gladwell (2002) described people, such as those who tweeted on the

final day of the tornado-outbreak case study, as Laggards. It is possible that attention spans are greater after a tornado outbreak when compared to a hazardous winter weather event, thus reaching the attention of Laggards. Houston et al. (2015) noted that social media posts often highlight damages after natural hazard events. Tweets from 5 March possibly discuss the outbreak's impact, especially in terms of damages and fatalities. Twenty-three residents of Lee County, Alabama, died in one particular tornado (Gore, 2019). Post-event tweets likely expressed condolences for persons directly affected. Regression did not well model the change in NNI over time during the tornado outbreak. This is perhaps related of the complex nature of the case study. The tornado-outbreak's spatial extent is of the synoptic scale, but an individual tornado is of the microscale. Cubic and inverse regression models had difficulty capturing such complex event, as indicated by low R^2 values for the tornado outbreak.

Population dynamics did not skew tornado-outbreak SDE mean center locations during the forecast and event period (28 February to 3 March) as was the case in the WSG study. Rather, proximity to reported tornado events seems to influence SDE mean center locations the most. Tornado-outbreak tweets also followed the event's forecast area fairly well from 28 February to 3 March. The 28 February and 1 March SDEs closely matched locations highlighted for Thunderstorm risk by Storm Prediction Center Convective Outlook maps (Dial, 2019; Mosier & Bentley, 2019). There were also clusters of tweets in southern Georgia and eastern Alabama on 3 March. This area almost perfectly aligns with the Enhanced risk seen on the Storm Prediction Center 3 March Convective Outlook map (Gleason & Karstens, 2019).

Tweets were not consistent with locations of Storm Prediction Center risk levels on 4 and 5 March (Gleason & Bentley, 2019; Mosier & Bentley, 2019). I hypothesize this is because

the event itself had already passed, so Convective Outlooks from 4 and 5 March are not related to my study. Tweets from 5 March displayed clustering around Atlanta, Georgia, suggesting an urban post-event bias in tweet. This shift towards urban tweet locations may be more prominent after high-impact events, because the news media reports the aftermath and draws the public's attention. People directly impacted by the tornado outbreak spend time cleaning and recuperating, but people farther away in cities may have the time to tweet. Also, tornado risk is less obvious during the forecast period than winter weather which can be predicted more accurately. The rise in post-event tweets might indicate that the tornado outbreak might have caught some people off-guard. The growth of SDE supports the hypothesis that interest was broader after the event. I would need additional research to confirm whether this happened after different tornado outbreaks, or was unique to my case study.

Though individuals across a whole region may prepare for a potential tornado outbreak, relatively few people are directly affected by any one tornado's path. Individual tornado events are spatially-constrained to a particular location. Tornadoes cause immense damage, but only to a specific location rather than the region at large. Communities more directly affected may more actively tweet about a weather event than ones farther away. Kryvasheyeu et al. (2016) observed this relationship in their analysis of tweets during Superstorm Sandy. Distance to hazardous weather events may explain why I observed fewer tweets near cities in the tornado-outbreak case study. On the other hand, WSG impacted cities such as Chicago and Washington D.C. The fact that WSG impacted cities could explain why I saw increased Twitter activity in these locations.

5.3 Tweet diffusion barriers during hazardous weather events

Some barriers described by Gould et al. (1969) possibly affected the diffusion patterns I observed. It may be more difficult to study Twitter diffusion patterns across remote and rural locations. Such places may experience technological, cultural, and physical barriers. Cellular data or a WIFI connection is needed to access Twitter, both of which may be lacking in a remote or rural location (technological and physical barrier). Residents of such locations may also not fit the typical Twitter user demographic (cultural barrier), and therefore may be at a disadvantage to receiving weather information via Twitter. Tornado outbreaks generate greater emotional response than winter weather. Tornadoes can effect entire communities and families, whereas winter weather generally passes with minimal impact. People may feel at-risk during a tornado-outbreak, and feel a greater sense of fear that may inhibit one from tweeting (psychological barrier). Tweets were clustered in urban locations throughout all WSG, but only during the finals days in the tornado-outbreak case study. From 28 February to 3 March, I observed tornado-outbreak tweets in locations directly impacted by the event. These people likely felt vulnerable to the severe weather, and expressed concerns on Twitter. People farther away from tornado locations possibly felt invincible, and therefore were less interested in tweeting until the news reported damages.

5.4 Uncertainties using regression to analyze spatiotemporal diffusion events

I used methods outlined by Lee et al. (2014) to compare the diffusion of tweets within two natural hazard case studies. Lee et al. (2014) examined how point data diffused through applications of nearest neighbor and regression, specifically identifying hierarchical, relocation, expansion, and contagious diffusion within point data. While these methods were successful for

Lee et al. (2014), I found fairly low R^2 values in all regressions except WSG's cubic model (R^2 of 0.836). Low R^2 values indicate regression models failed to depict data variance. WSG's inverse model had an R^2 value of 0.459 while R^2 values for the tornado outbreak's cubic and inverse models were 0.159 and 0.123, respectively. Although R^2 is unimpressive in the tornado outbreak, both case studies reported highest R^2 values with the cubic regression model.

Per Lee et al. (2014)'s findings, higher cubic regression R^2 values indicate that data exhibit hierarchical or relocation diffusion. However, I see aspects of all four diffusion classifications in each case study, which may be why a single model did not perform well. Hazardous weather information trickles through various "gate keepers" before it reaches the general public (hierarchical diffusion). Winter and tornadic weather information first passes through NOAA's Weather Prediction Center (located in College Park, Maryland) and Storm Prediction Center (located in Norman, Oklahoma), respectively. Weather Prediction Center and Storm Prediction Center staff members then share information to NOAA regional centers and Weather Forecast Offices (WFOs). The 126 WFOs nationwide distribute information within their local communities. Media outlets, government officials, and emergency managers then broadcast weather information to the general public. National NOAA offices, local WFOs, and community leaders act as "gate keepers." The public is only made aware of weather information after first passing through different sources. I also can identify evidence of expansion diffusion. Tweet counts generally grew in days preceding each case study event. Cognizance of WSG and the tornado outbreak "expanded" over time. Expansion can also be "contagious" (contagious diffusion) when users engage with weather information through their Twitter network (direct contact). Examples of engagement include actively writing and

disseminating weather information, or reading and absorbing weather content posted previously by another user. Direct contact, through the virtual space, exists in either case because Twitter users interact through a shared social media network. Tweets are, therefore, often contagious. I observed relocation diffusion broadly when I analyzed the change in SDE mean center locations. SDE mean centers "relocated," from one place to another, throughout the course of each event. I expected some relocation to be evident with WSG because of the system's progression west to east, which may be why NNI was best modeled with the cubic regression.

5.5 Limitations

The time required to filter tweets was one limitation to my thesis work. As a result of my manual filtering process and its limits, some tweets that passed through were irrelevant to my study. Tweets related to Mardi Gras is one example. Many of these tweets contained the tornado-outbreak keyword "hail," but were in reference to holiday festivities. The words "Mardi Gras" did not appear in Figures 3 or 4, which is why I did not eliminate these tweets from my study. Machine learning can be used in future studies to more efficiently filter tweets.

Sample sizes, and keywords used to acquire tweets were also limitations. Tornado-outbreak (WSG) tweets were acquired by inputting 11 (3) keywords into the API. The API initially returned 14,983 tornado outbreak and 25,836 WSG tweets. Upon filtering, 213 tornado outbreak and 2,881 WSG tweets remained. I therefore used only 1.42% $\left(\frac{213}{14,983}\right)$ tornado outbreak and 11.15% $\left(\frac{14,983}{25,836}\right)$ tweets in my data analysis. Tornado-outbreak tweets, acquired through the API, were noisy and largely off-topic. The keyword "basement" presented issues

which I determined through my attempt at filtering tweets through random samples. Instead of tweets mentioning “basement” in a protective-action sense such as, “I’m going to the basement because there are reports of a tornado near,” some tweets referred to The Basement (a bar in Atlanta, Georgia). I noticed similar issues with the keywords “hail” and “shelter,” when I tried to filter tweets. “Hail” tweets largely referenced Mardi Gras festivities while “shelter” tweets promoted animal shelters. I thought more API keywords would result in higher quality tweets that were more related to the tornado outbreak itself, but this was not the case. I suggest that researchers user fewer, more specific keywords to yields more relevant tweets from Twitter’s APIs. Low sample size may also be attributed to removing duplicate tweets, as part of my tweet cleaning scheme. It is possible that retweeted tweets were removed when eliminating duplicates. Expansion diffusion may be observed by analyzing retweets. For example, expansion diffusion would be observed if many people retweeted a celebrity’s tweet. The tornado-outbreak sample size would likely be larger if duplicate tweets (and consequently, retweets) were not eliminated. This would possibly alter tweet count, NNI, regression, and SDE results.

CHAPTER SIX

CONCLUSION

I researched how tweet diffusion varies spatiotemporally during two case studies. I refer to WSG as a "multi-day, synoptic" system because it occurred over multiple days and regions. I consider my second case study, a tornado outbreak, a regional event because it occurred in the Southeast. Technically, this event is also caused by synoptic-scale forcing, but the storms actually causing the tornadoes are mesoscale in nature. Though atmospheric conditions favored severe weather for days prior, I selected 3 March for my second case study because 80 tornadoes formed on this day. WSG covered more space (2.5k kilometers versus 1.7k kilometer) than the tornado-outbreak area. Thousands of people lost electricity during WSG and 23 people lost their lives during a particular tornado in central Alabama. Both hazardous weather events created a human impact; however, daily tweet counts indicate people remained interested in the tornado outbreak longer than WSG.

The tweets for this work were acquired by inputting keywords into Twitter's Streaming and Searching APIs. The APIs produced 25,836 WSG and 14,983 tornado-outbreak tweets. Filtering techniques reduced counts to 2,844 (WSG) and 213 (tornado outbreak) tweets. I first observed daily tweet counts. Pre-event tweets steadily rose in the tornado-outbreak case study, but dipped one day in WSG. The greatest number of tweets occurred during the peak of each hazardous weather event. Post-event, tweet counts dropped immediately following event peaks, but, tweets rebounded slightly on the final day of the tornado-outbreak case study.

I analyzed tweet diffusion in two different ways. First, I used methods outlined by Lee et al. (2014); specifically, I regressed NNI values over time using cubic and inverse regression models. My regression models generally performed poorly, especially in the tornado-outbreak case study, as indicated by low R^2 values. This finding suggests regression is not a suitable method for classifying hazardous weather tweet diffusion. High variance was observed over early hours in the regression analysis (as seen in Figures 8 and 9). This is possibly attributed to NNI calculations based on too few points. In other words, there is possibly a minimum number of NNI values that must be present in a system prior to using as the dependent variable in a regression model. A complicated pattern emerges, as tweets are shared in virtual space while weather systems move across physical space. The tornado-outbreak case study also exists over multiple scales. The entire event is of the synoptic scale, while an individual tornado is of the microscale. Simple regression models cannot account for complex phenomena that occur over multiple spaces and, as is the case in the tornado outbreak, multiple scales. Next, I analyzed the data using SDEs, and assessed how tweet mean center locations changed throughout the duration of the events. WSG tweets did not seem to follow the system's path. Concentrations were consistently highest in the Ohio River Valley and Mid-Atlantic, even though WSG formed over the Great Plains and Mid-West first. Tornado-outbreak SDE and tweets aligned with locations warned for severe convective weather by the Storm Prediction Center, rather than locations with high population densities. The final two days are an exception, as tweet counts rose due to broader interest in the overall event. My findings support the hypothesis that WSG demonstrated relocation diffusion because R^2 values were highest in the cubic regression model. Expansion diffusion was not observed in the tornado-outbreak case study, as originally

predicted. R^2 values were rather low, and the event's complicated nature was not well-modeled via regression.

My thesis, in combination with ongoing and future works in this area, will help the research community better understand when people become aware of natural hazard events and information. Multidisciplinary research rooted in psychology, sociology, and human dynamics can be used to approach the topics of weather awareness and spatiotemporal information diffusion from multiple angles. A thorough content analysis can be used for a more comprehensive analysis of tweet diffusion. Tweets often convey emotion, opinion, and tone—none of which were captured by my thesis because I only examined tweet counts and tweet locations. Had the APIs provided more relevant tweets, I would have looked into this further. Additional research examining tweet content may help researchers better understand when and how people become mindful of weather information shared on social media.

Future studies may also determine whether results are unique to my case studies, or if other weather events lead to similar findings. Specifically, it is unclear whether other tornadoes generate interest on Twitter after initial events pass, suggested by my tornado-outbreak case study. Future studies can use my methods to study tweet diffusion throughout different contemporary or historical case studies. Hu (2018) examined tweets counts before, during, and after Hurricane Irma (2017); however, methods used did not include NNI, regression, or SDE. Coastal NWS offices may take particular interest in studies analyzing hurricane-related tweet diffusion to gain insight of what community members are thinking and feeling during a hazardous weather event. Such valuable information can help community leaders better educate and inform the public during a potentially chaotic situation.

While WSG impacted multiple regions including the Rockies, Mid-West, Ohio River Valley, and Mid-Atlantic, I did not perform a comprehensive regional tweet analysis. There are likely different barriers as well as social, cultural, or climatological factors that may help explain tweet behavior across different regions. Mid-Westerners, for instance, may be more conditioned to winter weather than people in the Mid-Atlantic. The Southeast, though studied as one area in my tornado-outbreak case study, can likely break down into smaller sub-regions. Each sub-region may also have different factors that could influence one to tweet. For example, people who live in rural locations may have more personal experience with tornadoes than individuals who live in cities. Future research can determine the effect of regions on tweet patterns.

I noticed SDEs and Convective Outlook risk levels aligned fairly well in the tornado-outbreak case study. One additional future study may include quantitatively comparing the overlap of SDEs and Convective Outlook risk levels in future tornado outbreaks. While this was the original goal for my work, the sample size in my study limits my ability to draw these conclusions.

I believe my research is important for communities. Local NWS forecasters, emergency managers, media personnel, and government officials routinely share weather information with the public. These individuals can use my research and similar works to understand how weather information spreads on Twitter, and improve hazardous weather communication or education efforts. Community leaders can use the findings I presented to better understand human experience during hazardous weather events.

REFERENCES

@NWSSPC. (2019). “2:02pm CST #SPC Day1 Outlook Enhanced Risk: across parts of southern alabama, the far western florida panhandle, southwest and south-central georgia http://go.usa.gov/YWq5.” Retrieved from https://twitter.com/NWSSPC/status/1102299452285640706

@NWSWPC. (2019a). “#Snow is slowly tapering off over the Midwest and Ohio Valley but will continue into the evening over the interior Mid-Atlantic. https://wpc.ncep.noaa.gov/#page=frt.” Retrieved from https://twitter.com/NWSWPC/status/1084541815792025600

@NWSWPC. (2019b). “Moisture begins to increase by Friday ahead of a developing low pressure system in the Southern Plains http://go.usa.gov/cu3Wx.” Retrieved from https://twitter.com/NWSWPC/status/1082971559420194816

@weatherchannel. (2019a). “A late-week snowmaker could impacts parts of the South, Midwest and East: https://wxch.nl/2FdJ2vd.” Retrieved from https://twitter.com/weatherchannel/status/1082428419349139458

@weatherchannel. (2019b). “After the weekend’s #WinterStormGia, another snowstorm is looking increasingly likely Friday -> next weekend in the Plains, Midwest and Northeast. https://wxch.nl/2HaZrlZ.” Retrieved from https://twitter.com/weatherchannel/status/1084782682532401152

@weatherchannel. (2019c). “BREAKING: #WinterStormScott has been named and will bring heavy #snow from the Rockies into the Northeast with a #severe storm threat across parts of the South. We are LIVE now with all the details so you can plan your weekend.” Retrieved from https://twitter.com/weatherchannel/status/1101604582625632256

@weatherchannel. (2019d). “Our latest thinking on Saturday’s #severe weather potential in the South, and whether Sunday’s threat is anything to worry about for areas recovering from last Sunday’s outbreak. https://wxch.nl/2UoBytl.” Retrieved from https://twitter.com/weatherchannel/status/1103989239254052865

@weatherchannel. (2019e). “See the before-and-after images of the damage caused by Sunday’s EF4 #tornado in Alabama: https://wxch.nl/2HjlxRx.” Retrieved from https://twitter.com/weatherchannel/status/1102969827864571904

@weatherunderground. (2019). “Sunday brings another chance for #severe storms across the South as our Coast-to-Coast storm brings #snow to the Plains & into the Northeast. Have a question on these stories? Tweet us using #WUTV & we’ll answer yours on the show!”. Retrieved from https://twitter.com/wunderground/status/1101266143627628544

About different types of Tweets. (2020). Retrieved June 24, 2020, from https://help.twitter.com/en/using-twitter/types-of-tweets

Access tokens. (2019). Retrieved September 8, 2019, from https://developer.twitter.com/en/docs/basics/authentication/guides/access-tokens.html

Acharya, B. K., Cao, C., Xu, M., Chen, W., & Pandit, S. (2018). Spatiotemporal Distribution and Geospatial Diffusion Patterns of 2013 Dengue Outbreak in Jhapa District, Nepal. *Asia-Pacific Journal of Public Health*, *30*(4), 396–405. https://doi.org/10.1177/1010539518769809

Advanced Search. (2020). Retrieved July 29, 2020, from https://twitter.com/search-advanced?lang=en

Alexander, D. E. (2014). Social Media in Disaster Risk Reduction and Crisis Management. *Science and Engineering Ethics*, *20*(3), 717–733. https://doi.org/10.1007/s11948-013-9502-z

Altay, N., & Pal, R. (2014). Information diffusion among agents: Implications for humanitarian operations. *Production and Operations Management*, *23*(6), 1015–1027. https://doi.org/10.1111/poms.12102

Bahk, C. Y., Cumming, M., Paushter, L., Madoff, L. C., Thomson, A., & Brownstein, J. S. (2016). Publicly Available Online Tool Facilitates Real-Time Monitoring Of Vaccine Conversations And Sentiments. *Health Affairs (Project Hope)*, *35*(2), 341–347. https://doi.org/10.1377/hlthaff.2015.1092

Beauregard-Smiths Station Tornado - March 3, 2019. (n.d.). Retrieved from https://www.weather.gov/bmx/event_03032019beauregard

Bohannon, J. (2016). Twitter can predict hurricane damage as well as emergency agencies. *Science*. https://doi.org/10.1126/science.aaf4182

Chew, C., & Eysenbach, G. (2010). Pandemics in the age of Twitter: Content analysis of tweets during the 2009 H1N1 outbreak. *PLoS ONE*, *5*(11), 1–13. https://doi.org/10.1371/journal.pone.0014118

Clark, P. J., & Evans, F. C. (1954). Distance to Nearest Neighbor as a Measure of Spatial Relationships in Populations. *Ecology*, *35*(4), 445–453.

Cliff, A., Pred, A., & Hagerstrand, T. (1992). Classics in Human Geography Revisited. *Progress in Human Geography*, *16*(4), 541–544.

Cooper, G. P., Yeager, V., Burkle, F. M., Subbarao, I., Jr., & Subbarao, I. (2015a). Twitter as a Potential Disaster Risk Reduction Tool. Part II: Descriptive Analysis of Identified Twitter Activity during the 2013 Hattiesburg F4 Tornado. *PLoS Currents*, *7*. https://doi.org/10.1371/currents.dis.f2e5b9e979af6174d2f97c1f0349be5c

Cooper, G. P., Yeager, V., Burkle, F. M., Subbarao, I., Jr., & Subbarao, I. (2015b). Twitter as a Potential Disaster Risk Reduction Tool. Part III: Evaluating Variables that Promoted Regional Twitter Use for At-risk Populations During the 2013 Hattiesburg F4 Tornado. *PLoS Currents*, *7*. https://doi.org/10.1371/currents.dis.b305fe1b479528fda724c6f84f546471

Counting characters. (2020). Retrieved June 24, 2020, from https://developer.twitter.com/en/docs/basics/counting-characters

DeCaria, A. (n.d.). ESCI 241-Meteorology Lesson 13 Small Scale Circulations. Retrieved July 25, 2020, from http://snowball.millersville.edu/~adecaria/ESCI241/esci241_lesson13_circulation_I.pdf

Di Liberto, T. (2019). Early March 2019 tornado outbreak ravages parts of U.S. Southeast. Retrieved May 29, 2019, from https://www.climate.gov/news-features/event-tracker/early-march-2019-tornado-outbreak-ravages-parts-us-southeast

Dial. (2019). Storm Prediction Center Mar 1, 2019 0600 UTC Day 1 Convective Outlook. Retrieved October 5, 2020, from https://www.spc.noaa.gov/products/outlook/archive/2019/day1otlk_20190301_1200.html

Evans, J. (n.d.). nni function. Retrieved September 11, 2020, from https://www.rdocumentation.org/packages/spatialEco/versions/1.3-2/topics/nni

Galarnyk, M. (2018). Explaining the 68-95-99.7 rule for a Normal Distribution. Retrieved July 15, 2020, from https://towardsdatascience.com/understanding-the-68-95-99-7-rule-for-a-normal-distribution-b7b7cbf760c2

Gladwell, M. (2002). *The Tipping Point How Little Things Can Make a Big Difference*. New York: Back Bay Books.

Gleason, & Bentley. (2019). Storm Prediction Center Mar 4, 2019 0600 UTC Day 1 Convective Outlook. Retrieved October 5, 2020, from https://www.spc.noaa.gov/products/outlook/archive/2019/day1otlk_20190304_1200.html

Gleason, & Karstens. (2019). Storm Prediction Center Mar 3, 2019 0600 UTC Day 1 Convective Outlook. Retrieved October 5, 2020, from https://www.spc.noaa.gov/products/outlook/archive/2019/day1otlk_20190303_1200.html

Gore, L. (2019). Alabama tornado 2019: 23 killed, including 3 children, more missing; What we know today - al.com. Retrieved September 28, 2020, from https://www.al.com/news/2019/03/alabama-tornado-2019-23-killed-including-3-children-more-missing-what-we-know-today.html

Gould, P. R. (1969). *Spatial Diffusion, Resource Paper No. 4*. *Spatial Diffusion*. Washington, D.C. https://doi.org/10.1016/0143-6228(89)90060-x

Gregory, T., & Patuelli, R. (2015). Demographic ageing and the polarization of regions—an exploratory space–time analysis. *Environment and Planning A*, *47*(5), 1192–1210. https://doi.org/10.1177/0308518X15592329

Hagerstrand, T. (1967). *Innovation diffusion as a spatial process*. (A. Pred & G. Haag, Eds.). Chicago: University of Chicago Press.

Hayes, A. (n.d.). broom package. Retrieved September 10, 2020, from https://www.rdocumentation.org/packages/broom/versions/0.7.0

Hayes, J. L. (2012). The Historic Tornadoes of April 2011. *National Oceanic and Atmospheric Administration Service Asessment*, (April).

Hoekstra, S., Klockow, K., Riley, R., Brotzge, J., Brooks, H., & Erickson, S. (2011). A Preliminary Look at the Social Perspective of Warn-on-Forecast: Preferred Tornado Warning Lead Time and the General Public's Perceptions of Weather Risks. *Weather, Climate, and Society*, *3*(2), 128–140. https://doi.org/10.1175/2011wcas1076.1

Hoffman, C. (2018). What Is an API? Retrieved June 28, 2020, from https://www.howtogeek.com/343877/what-is-an-api/

Holmes, S. (2004). Monte Carlo. Retrieved July 7, 2020, from http://statweb.stanford.edu/~susan/courses/s208/node14.html

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

How Directional Distribution (Standard Deviational Ellipse) works. (n.d.). Retrieved July 15, 2020, from https://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/h-how-directional-distribution-standard-deviationa.htm

Hu, Y. (2018). Geo-text data and data-driven geospatial semantics. *Geography Compass*, *12*(11). https://doi.org/10.1111/gec3.12404

Inhibit Interpretation/Conversion of Objects. (n.d.). Retrieved September 10, 2020, from https://stat.ethz.ch/R-manual/R-devel/library/base/html/AsIs.html

Iowa State University. (2019). IEM :: NWS Watch Warning Archive Download. Retrieved April 9, 2019, from https://mesonet.agron.iastate.edu/request/gis/watchwarn.phtml

Jackson, B. (2019). *Central Rockies to Mid-Atlantic Winter Storm (11 - 14 January, 2019)*. Retrieved from https://www.wpc.ncep.noaa.gov/storm_summaries/event_reviews.php?YYYYMMDD=20190112

Kim, Y., Nordgren, R., & Emery, S. (2020). The story of goldilocks and three twitter's APIs: A pilot study on twitter data sources and disclosure. *International Journal of Environmental Research and Public Health*, *17*(3), 1–15. https://doi.org/10.3390/ijerph17030864

Kryvasheyeu, Y., Chen, H., Obradovich, N., Moro, E., Hentenryck, P. Van, Fowler, J., & Cebrian, M. (2016). Rapid assessment of disaster damage using social media activity. *Science Advance*, (March), 1–12. https://doi.org/10.1126/sciadv.1500779

Le Texier, M., & Caruso, G. (2017). Assessing geographical effects in spatial diffusion processes: The case of euro coins. *Computers, Environment and Urban Systems*, *61*, 81–93. https://doi.org/10.1016/j.compenvurbsys.2016.08.003

Lee, J., Lay, J.-G., Chin, W. C. B., Chi, Y.-L., & Hsueh, Y.-H. (2014). An Experiment to Model Spatial Diffusion Process with Nearest Neighbor Analysis and Regression Estimation. *International Journal of Applied Geospatial Research*, *5*(1), 1–15. https://doi.org/10.4018/ijagr.2014010101

Lee, J., Li, S., Wang, S., Wang, J., & Li, J. (2020). Spatio-Temporal Nearest Neighbor Index for Measuring Space-Time Clustering among Geographic Events. *Papers in Applied Geography*, *0*(0), 1–14. https://doi.org/10.1080/23754931.2020.1810112

Ma, Z. (2011). Hägerstrand time-geography revisited from current perspective. *2011 International Conference on Multimedia Technology, ICMT 2011*, (128), 5407–5410. https://doi.org/10.1109/ICMT.2011.6002703

Malik, M. M., Lamba, H., & Nakos, C. (2015). *Population Bias in Geotagged Tweets*. (D. Ruths & J. Pfeffer, Eds.), *Standards and Practices in Large-Scale Social Media Research: Papers from the 2015 ICWSM Workshop*. Palo Alto: The Association for the Advancement of Artifical Intelligence.

March 3rd, 2019 Tornado Outbreak. (2019). Retrieved July 25, 2020, from https://www.weather.gov/cae/Tornado_20190303

Martín, Y., Li, Z., & Cutter, S. L. (2017). Leveraging Twitter to gauge evacuation compliance: Spatiotemporal analysis of Hurricane Matthew. *PLoS ONE*, *12*(7), 1–22. https://doi.org/10.1371/journal.pone.0181701

Mason, L. R., Ellis, K. N., Winchester, B., Schexnayder, S., Mason, L. R., Ellis, K. N., … Schexnayder, S. (2018). Tornado Warnings at Night: Who Gets the Message? *Weather, Climate, and Society*, *10*(3), 561–568. https://doi.org/10.1175/WCAS-D-17-0114.1

Morss, R. E., Demuth, J. L., Lazrus, H., Palen, L., Barton, C. M., Davis, C. A., … Watts, J. (2017). Hazardous weather prediction and communication in the modern information environment. *Bulletin of the American Meteorological Society*, *98*(12), 2653–2674. https://doi.org/10.1175/BAMS-D-16-0058.1

Morstatter, F., Pfeffer, J., & Liu, H. (2014). When is it biased? Assessing the representativeness of twitter's streaming API. *WWW 2014 Companion - Proceedings of the 23rd International Conference on World Wide Web*, 555–556. https://doi.org/10.1145/2567948.2576952

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the Sample Good Enough ? Comparing Data from Twitter ' s Streaming API with Twitter ' s Firehose. *Association for the Advancement of Artificial Intelligence*, 400–408.

Mosier, & Bentley. (2019). Storm Prediction Center Mar 5, 2019 0600 UTC Day 1 Convective Outlook. Retrieved October 5, 2020, from https://www.spc.noaa.gov/products/outlook/archive/2019/day1otlk_20190305_1200.html

Nguyen, Q. C., McCullough, M., Meng, H. W., Paul, D., Li, D., Kath, S., … Li, F. (2017). Geotagged US tweets as predictors of county-level health outcomes, 2015-2016. *American Journal of Public Health*, *107*(11), 1776–1782. https://doi.org/10.2105/AJPH.2017.303993

Palen, L., Vieweg, S., Liu, S. B., & Hughes, A. L. (2009). Crisis in a Networked World. *Social Science Computer Review*, *27*(4), 467–480. https://doi.org/10.1177/0894439309332302

Paniagua, J., & Sapena, J. (2014). Business performance and social media: Love or hate? *Business Horizons*, *57*(6), 719–728. https://doi.org/10.1016/j.bushor.2014.07.005

Perrin, A., & Anderson, M. (2019). Social media usage in the U.S. in 2019 | Pew Research Center. Retrieved June 24, 2020, from https://www.pewresearch.org/fact-tank/2019/04/10/share-of-u-s-adults-using-social-media-including-facebook-is-mostly-unchanged-since-2018/

Principles of Epidemiology: Glossary. (2014). Retrieved October 17, 2020, from https://www.cdc.gov/csels/dsepd/ss1978/glossary.html

Qu, Y., Wu, P. F., & Wang, X. (2009). Online Community Response to Major Disaster: A Study of Tianya Forum in the 2008 Sichuan Earthquake. In *2009 42nd Hawaii International Conference on System Sciences* (pp. 1–11). IEEE. https://doi.org/10.1109/HICSS.2009.330

R-core R-core@R-project.org. (n.d.). poly function. Retrieved September 10, 2020, from https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/poly

R Language Definition. (2020), *4.0.3*, 10. https://doi.org/10.1002/9781119197515.app4

Rey, S. J., Murray, A. T., & Anselin, L. (2011). Visualizing Regional Income Distribution Dynamics. *Letters in Spatial and Resource Sciences*, *4*, 81–90. https://doi.org/https://doi.org/10.1007/s12076-010-0048-2

Ripberger, J. T., Jenkins-Smith, H. C., Silva, C. L., Carlson, D. E., Henderson, M., Ripberger, J. T., … Henderson, M. (2014). Social Media and Severe Weather: Do Tweets Provide a Valid Indicator of Public Attention to Severe Weather Risk Communication? *Weather, Climate, and Society*, *6*(4), 520–530. https://doi.org/10.1175/WCAS-D-13-00028.1

Santorelli. (2019a). 1200Z Surface Analysis for Sat Mar 02 2019. NOAA WPC, NHC, OPC. Retrieved from https://www.wpc.ncep.noaa.gov/archives/sfc/2019/ussatsfc2019030212.gif

Santorelli. (2019b). 1200Z Surface Analysis for Sun Mar 03 2019. NOAA WPC, NHC, OPC. Retrieved from https://www.wpc.ncep.noaa.gov/archives/sfc/2019/ussatsfc2019030312.gif

Schramm, M. E., Trainor, K. J., Shanker, M., & Hu, M. Y. (2010). An agent-based diffusion model with consumer and brand agents. *Decision Support Systems*, *50*(1), 234–242. https://doi.org/10.1016/j.dss.2010.08.004

Search Tweets Standard search API. (2020). Retrieved June 28, 2020, from https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets

Severe Weather Event Summaries. (2018). Retrieved July 29, 2020, from https://www.spc.noaa.gov/climo/online/

Shaw, S. L., & Sui, D. (2018). Editorial: GIScience for human dynamics research in a changing world. *Transactions in GIS*, *22*(4), 891–899. https://doi.org/10.1111/tgis.12474

Stilo, G., Velardi, P., Tozzi, A. E., & Gesualdo, F. (2014). Predicting Flu Epidemics Using Twitter and Historical Data (pp. 164–177). Springer, Cham. https://doi.org/10.1007/978-3-319-09891-3_16

Storm Prediction Center. (2019). *SPC Filtered Storm Reports for 03/03/19*. NOAA's National Weather Service. Retrieved from https://www.spc.noaa.gov/climo/reports/190303_rpts.html

Storm Prediction Center Storm Reports for 3/3/19. (2019). Retrieved October 17, 2020, from https://www.spc.noaa.gov/climo/reports/190303_rpts.html

Sullivan, L., & LaMorte, W. W. (2016). Variance and Standard Deviation. Retrieved July 15, 2020, from https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_SummarizingData/BS704_SummarizingData6.html

Team, R. C. (2020). *R Language Definition*.

The Weather Channel Announces Winter Storm Names for 2019-2020 Season. (2019). Retrieved July 29, 2020, from http://weathergroup.com/newsroom/2019/10/03/weather-channel-announces-winter-storm-names-2019-2020-season

Tornadoes of March 3, 2019 Event Summary for Central Alabama. (n.d.). Retrieved July 25, 2020, from https://www.weather.gov/bmx/event_03032019

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

Vaught, S. E., & Castagno, A. E. (2008). "I don't think I'm a racist": Critical Race Theory, teacher attitudes, and structural racism. *Race Ethnicity and Education*, *11*(2), 95–113. https://doi.org/10.1080/13613320802110217

Wang, Y., Callan, J., & Zheng, B. (2015). Should we use the sample? Analyzing datasets sampled from Twitter's stream API. *ACM Transactions on the Web*, *9*(3). https://doi.org/10.1145/2746366

Watch/Warning/Advisory Definitions. (n.d.). Retrieved September 23, 2019, from https://www.weather.gov/lwx/WarningsDefined

Welbers, K., & Opgenhaffen, M. (2018). Social media gatekeeping: An analysis of the gatekeeping influence of newspapers' public Facebook pages. *New Media and Society*, *20*(12), 4728–4747. https://doi.org/10.1177/1461444818784302

Wilkins, D. E., & Ford, R. L. (2007). Nearest neighbor methods applied to dune field organization: The Coral Pink Sand Dunes, Kane County, Utah, USA. *Geomorphology*, *83*(1–2), 48–57. https://doi.org/10.1016/j.geomorph.2006.06.009

Wolff, J., & Zalta, E. N. (Ed. . (2017). Karl Marx. In *Stanford Encyclopedia of Philosophy* (Winter 201). Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/entries/marx/#Bib

Zornoza, J. (2019). Downloading Data From Twitter Using the Streaming API. Retrieved June 24, 2020, from https://medium.com/@jaimezornoza/downloading-data-from-twitter-using-the-streaming-api-3ac6766ba96c

VITA

Alyssa Cannistraci finds science empowering and interesting. She earned her Bachelor of Science degree in meteorology from Millersville University. Alyssa received the National Oceanic and Atmospheric (NOAA) Ernest F. Hollings Undergraduate Scholarship, and completed an internship at the National Weather Center in Norman, Oklahoma. Through this internship, she realized a social science component exists within applied atmospheric research. She considered different graduate school options, and discovered geography at the University of Tennessee, Knoxville. She learned the discipline excellently combines social and physical science. Alyssa volunteered as a Social Media Ambassador and Website Coordinator for the Department of Geography. She is not ruling out a future PhD, possibly in communications or sociology; however, for now, she is excited to start a meaningful career. Alyssa seeks a position in science communication, education, and outreach.