Integrating Social Media in the Development of a Special Event Population Dynamics Model

Kelly Michelle Sims

University of Tennessee - Knoxville, ksims@vols.utk.edu

Recommended Citation
https://trace.tennessee.edu/utk_gradthes/3182
To the Graduate Council:

I am submitting herewith a thesis written by Kelly Michelle Sims entitled "Integrating Social Media in the Development of a Special Event Population Dynamics Model." I have examined the final electronic copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Geography.

Budhendra L. Bhaduri, Major Professor

We have read this thesis and recommend its acceptance:

Bruce A. Ralston, Hyun Kim

Accepted for the Council:

Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)
Integrating Social Media in the Development of a Special Event
Population Dynamics Model

A Thesis Presented for the
Master of Science
Degree
The University of Tennessee, Knoxville

Kelly Michelle Sims
December 2014
ACKNOWLEDGEMENTS

My deepest appreciation is owed to the following persons for their immense support and guidance provided to me during this research:

**Dr. Budhendra Bhaduri** (advisor), for his continuous encouragement and professional advice throughout this entire process. Without his wisdom, instructions, and true care for my success, this research would be non-existent.

**Dr. Bruce Ralston** (committee member), for his invaluable teaching and commitment to my education. His relentless confidence pushed me to improve my work wherever possible.

**Dr. Hyun Kim** (committee member), for his constructive suggestions and unwavering assistance to this research.

**U.S. Government**, for the financial support that allowed me to study at the university, as well as work simultaneously as a graduate research associate for Oak Ridge National Laboratory.

**Dr. Gautam Thakur & David Resseguie (ORNL)**, for their time and service in the collection of the social media data.

**Amy Rose** (ORNL mentor), for her constant support as a mentor, friend, and fellow graduate student also trying to finish her dissertation. Her numerous pep-talks reminded me that there was in fact, a light at the end of the thesis tunnel.

**Eric Weber, Jessica Moehl, Linda Sylvester, Melanie Phillips, Jake McKee, & Jason Kaufman** (ORNL), for all of their advice and recommendations. From the countless questions answered, to the suggested cautions of “what not to do,” these individuals were essential to my success as a graduate student.

**Thomas & Deborah Sims** (parents), for their enormous amount of faith in me. Without their love and motivation, I would not have accomplished many of the milestones in my life.

To the rest of my family, friends, and ORNL co-workers, thank you for being there when I needed to talk, vent, or laugh. Through the good and the bad, you helped me through it, and I am forever grateful.

Thank you all!
ABSTRACT

With society’s increasing participation in social media, scientists now have access to new sources of data that reflect our daily activities in space and in time. Such data are plentiful and, more notably, at an unprecedented granular level. The ability for users to capture and express their geolocation through their phones’ global positioning system (GPS) or through a particular location’s hashtag or Facebook Page provides a great opportunity for modeling spatiotemporal population dynamics. High resolution population models and databases for episodic special events can be extremely useful for enhancing emergency management and response. This research assesses the feasibility of improving a special event population distribution and dynamics model, namely Oak Ridge National Laboratory’s LandScan USA, using data from social media. Specifically, analysis is across a 24 hour period for a number of football game days associated with the University of Tennessee, Knoxville during the 2013-2014 season. Data from two popular social media platforms, namely Twitter and Facebook, were used to analyze possible patterns of population distributions around the university’s football stadium. Spatial autocorrelation was measured and calculated using Global Moran’s I and the Local Indicator of Spatial Association (LISA) test to support and build confidence of the tweet and check-in data. Overall, data from social media were found to be beneficial for improving high-resolution population distribution datasets, such as LandScan USA. However, long term collection and analysis of social media data are necessary for ensuring sustainability and predictive capacity of such data in modeling near real-time population dynamics for special events.
# TABLE OF CONTENTS

CHAPTER I INTRODUCTION .............................................................................................................1  
  Population Dynamics .......................................................................................................................2  
  Special Events .................................................................................................................................3  
  Social Media ..................................................................................................................................4  
CHAPTER II BACKGROUND .................................................................................................................7  
  Cyberspace & Neogeography ..........................................................................................................7  
    Volunteered Geographic Information ............................................................................................8  
    Twitter .........................................................................................................................................9  
    Facebook .....................................................................................................................................10  
    Limitations of Social Media .........................................................................................................11  
    Applications of Social Media Data ...............................................................................................12  
  Nature of the Problem ....................................................................................................................13  
  Objective .......................................................................................................................................15  
CHAPTER III METHODS ....................................................................................................................17  
  Streaming Collection ......................................................................................................................17  
    Collected Data ............................................................................................................................19  
  Verification & Validation .................................................................................................................20  
    Limitations of Twitter ..................................................................................................................22  
    Limitations of Facebook Check-in Data .......................................................................................23  
  Standardization & Testing ..............................................................................................................24  
  Model for Future Use .....................................................................................................................26  
CHAPTER IV RESULTS & DISCUSSION ...........................................................................................28  
  Tweets and Check-ins ....................................................................................................................28  
    Facebook Results .......................................................................................................................31  
    Popularity of Tweeted Terms .......................................................................................................32  
  Spatial Significance .......................................................................................................................34  
    Global Moran’s I ..........................................................................................................................34  
    Local Clustering ..........................................................................................................................39  
  Special Events Model ....................................................................................................................40  
    “Before The Game” Model .........................................................................................................41  
    “During The Game” Model ..........................................................................................................44  
  Discussion ......................................................................................................................................48  
CHAPTER V CONCLUSION ................................................................................................................52  
LIST OF REFERENCES .......................................................................................................................54  
APPENDIX ........................................................................................................................................58  
VITA ................................................................................................................................................101
LIST OF TABLES

Table 1. (a,b,c). Twitter’s Heat Map of the ‘Top 7’ terms in an aggregate 24 hour window for multiple game type scenarios. .................................................................................................................33
Table 2. Global Moran’s I summary of each search term’s associated tweets, as well as the combined total of all tweets. .................................................................................................................36
Table 3. Global Moran’s I report of spatial autocorrelation for the combined Top 7 Terms at multiple scales from Neyland Stadium. .................................................................................................................37
Table 4. Global Moran’s I report of spatial autocorrelation for the combined Top 7 Terms used at Home or Away games, for multiple scales.................................................................................................................38
LIST OF FIGURES

Figure 1. Two models representing the 2012 LandScan USA Night (left) and Day (right) distributions around Atlanta, Georgia’s downtown area. .................................................................3
Figure 2. An example of ex post facto modeling techniques of a San Francisco Special Event Model 2 hours before game (left) and during game (right) when attendance is assumed at its maximum .................................................................14
Figure 3. 2013 Season outlook for the University of Tennessee’s football game schedule. ........19
Figure 4. The term UTK, meaning ‘for’ in the Indonesian language, provides an example of false-positive data in Tweets collected during University of Tennessee football games. .....23
Figure 5. Neyland Stadium’s Facebook check-in activity for all 5 home games, two hours prior to, and 30 minutes following kickoff .................................................................24
Figure 6. Total number of tweets for each home game compared to that respected home game’s attendance .................................................................29
Figure 7. Total number of check-ins for each home game compared to that respected home game’s attendance .................................................................30
Figure 8. A seasonal outlook of each game day’s check-in activity 12 hours before and 12 hours after kickoff .................................................................31
Figure 9. Local Moran’s I output raster of COTypes 1.5 miles surrounding Neyland Stadium .................................................................40
Figures 10(a&b). Using ArcMap’s Kernel Density Tool, these figures represent the weighted input rasters of tweets (10a) and check-ins (10b) for data collected before 2 hours prior to kick off and after 3 hours following kickoff .................................................................43
Figures 11(a&b). Using ArcMap’s Kernel Density Tool, these figures represent the weighted input rasters of tweets (11a) and check-ins (11b) based on those data collected from 2 hours before to 3 hours following kickoff .................................................................45
Figure 12. LandScan USA nighttime baseline dataset at 1.5 miles surrounding Neyland Stadium at the University of Tennessee, Knoxville .................................................................46
Figure 13. The Special Event population distribution models for the University of Tennessee’s football game days for 2.5 hours before kickoff (13a) and 30 minutes after kickoff (13b). .................................................................47
CHAPTER I
INTRODUCTION

The ever growing phenomenon of microblogging on social media provides researchers from many disciplines access to a new form of data that in turn allows new ways to interpret or model human activities in space and time. From ‘posting’ one’s current activities, to ‘checking-in’ to a local hotspot, user-generated content (UGC) offers detailed spatio-temporal information of human dynamics. Since such information is produced by a general population, it is not traditionally considered authoritative, but is voluntary information about the world around us meant to be received by others, usually through the Internet (Krumm, Davies, and Narayanaswami 2008). The intrigue from the sciences is to capture this plethora of material being discussed and shared on public platforms, and quantify it to better represent the dynamics of the continually changing world. Constant updates of current activities are widespread and endless, and most importantly, offer relatively easy access to data. Specifically for geographers, social media has many scales and locational attributes that may prove advantageous to population modeling. From the metadata to actual UGC, social media has the ability to offer more accurate representations of population dynamics for emergency management and response.

The potential benefits of integrating social media into population distribution models are explored in this research with explicit focus on episodic populations which have large attendance numbers and a significant presence on social networking services; in particular, the episodic population for this study is game day college football fans from the University of Tennessee (UT), Knoxville. This population was chosen for their reoccurring, yet episodic nature throughout a football season, and because of their large presence on social media’s two most popular platforms, Twitter and Facebook. Using tweets and check-ins, this research assesses if there are any measurable and/or predictable structures to game day social media dynamics, as well as any population density dynamics. By examining this particular event, local knowledge is implemented to control measurability and dependability for this type of UGC, thus minimizing the gaps of unfamiliarity. To ensure this approach could eventually be replicated for similar events, the dynamics of populations not normally accounted for in current population distribution models are analyzed using tweets and check-ins for the 2013-2014 football season.
Population Dynamics

Population dynamics conveys spatial and temporal attributes to depict where people are located, as well as model the patterns in which they move. For Geographic Information Systems (GIS) Analysts concerned with emergency and disaster management, improving population dynamics is crucial. Because natural and technological disasters can happen without prior warning, and can disrupt large populations, it is essential to have an understanding of where people are at any time of the day for effective crisis response. However, developing high resolution spatiotemporal population distribution data is a challenging task. Most efforts at modeling temporal populations lack day-time or diurnal distributions (Kobayashi et al. 2011) making proper emergency analysis only useful at night. Others lack episodic and tourist populations (Bhaduri 2008, Jochem et al. 2012) making only residential populations properly accounted for. Furthermore, censuses simply provide population counts or estimates at country or province levels (Dobson et al. 2000) and are often times decennial, creating large density outputs or out of date representations. Therefore, publicly available data feeds of social networks present an opportunity to overcome these limitations (Bukhari et al. 2012) for a near real-time representation of population patterns. By coupling social media with geographic information systems (GIS), additional improvements and refinements can be made for population distribution models and databases.

At Oak Ridge National Laboratory (ORNL), a high-resolution population distribution and dynamics model, known as LandScan USA (Bhaduri et al. 2007), has been developed for the entire U.S. depicting a nighttime (residential) and daytime population (Figure 1). At a spatial resolution of 3 arc-second (~90m cells), LandScan USA has become the community standard for agencies interested in homeland security, socio-economic studies, and infrastructure analysis (Bhaduri et al. 2007, Bhaduri 2008, Brass 2008). For the nighttime model, census block counts are spatially disaggregated using a dasymetric modeling approach. The daytime model utilizes specialized input datasets, such as worker mobility and student enrollments, for enhancing the distribution of where people are during the day (Bhaduri et al. 2007). Currently, the dataset disseminates populations using ancillary spatial data (i.e. infrastructure) and socio-economic data (i.e. journey to work) for an output of a normal weekday and night. However, present endeavors have recently explored ways to incorporate tourist and episodic populations (Jochem et al. 2013)
for a more accurate representation of populations. As such, social media can offer an additional avenue to further refine these findings of traditionally unaccounted groups, and is fundamentally where this research is of service. Of course, modeling at a national scale is labor intensive, as well as computationally demanding, so focus was narrowed to one highly populated event. This event will be the case study for all highly populated episodic events which are termed in this paper, “Special Events.”

**Special Events**

Merriam-Webster (2013) defines the word ‘event’ in a generic sense of “something that happens (or might happen).” In respect to this research, this word has a more specific meaning based on a particular occasion that is highly attended and popular to the general population. It also alters the normal distribution of an area by increasing the number of not normally accounted for populations. As such, the term is enhanced to ‘Special Event,’ and redefined as:

- Anomalous population density of a place resulting from a public gathering involving a significant influx of population to a particular location for a distinct period of time.

Examples of Special Events can include nationally publicized events (i.e. Democratic and Republican National Conventions), sports (i.e. Super Bowl), concerts (i.e. Bonaroo), fairs/festivals (i.e. Mardi Gras), etc. By narrowing the definition of the traditional term, emphasis is centered on these before-mentioned attractive events which are, most likely, well discussed on social media. These events should have trending dialogues which should give suitable insight.
into understanding the spatiotemporal dynamics of those attending these special events, particularly if those data are georeferenced. Additionally, Special Events with greater significance to the general public could have an elevated potential of population displacement, so it is imperative to study the social media data in the social context. However, because there are multiple special events that occur nationwide, this paper will focus on one, specific, special set of events: the University of Tennessee’s football games, held at Neyland Stadium and away, during the 2013-2014 season.

Like most collegiate sporting events, UT’s football program is a well-established entity of the university that captivates a tremendous following from the general public almost every Saturday during the Fall months. With 12 games scheduled to play either at home, in Neyland Stadium, or on the road against another university, these football games are a great example of a special event. In 2013, around 100,000 fans attended each of the seven games in Neyland Stadium, which created an extremely large influx of population to a fairly small area. Plus, the fact that these types of games offer many pre and post-game activities, enormous amounts of fans spend a significant amount of time on campus and the surrounding areas which can change the density and distributions of an area not normally associated with such large populations for the remainder of the week/year. Therefore, this research is necessary to offer additional models that represent special event populations in case emergency management analysis is required.

Social Media

Currently, the World’s two leading social media platforms are Twitter, with 230 million monthly active users\(^1\), and Facebook, with 874 million monthly active users\(^2\). The enormous popularity of these communication services has individuals and groups of all types engaging in information exchange all day. Known as microblogging, users share through social networks excerpts of their whereabouts, opinions, beliefs, activities, etc., for the sole purpose of broadcasting one’s life to interested parties (Grace, Dejin, Boyd 2010, Java et al. 2007). The content of the material is

boundless, but most importantly, voluntary. This proves worthwhile for application program interfaces (APIs) which can pragmatically mine a continuous sample of tweets (Twitter) and check-ins (Facebook, Places) to give insight into the world around us. From the most detailed of ways (i.e. explaining a personal itinerary of one’s vacation plans) to the most unimportant of ways (i.e. sharing a photo of one’s dinner), this chaotic and unsystematic availability of user information is what has provoked a fascinating relationship between scientists and Big Data.

With an abundance of knowledge at every cyber turn, the challenge becomes quickly organizing these large datasets for faster analysis to assist in swift emergency response.

Big Data is a term that currently lacks a unified definition and multiple descriptions and explanations are constantly being proposed by various user communities. Captivating many different sciences interested in massive collections of any quantifiable data, Big Data can range across a variety of subject matter from structured historical weather recordings to national medical reports to rather unstructured social media feeds. It is material either highly organized or entirely heterogeneous, which is often difficult to manage and process with traditional analytical tools (Manovich 2011). Nevertheless, Big Data fundamentally provides statistical insight into diverse and complex sciences. And though its definition is evolving, Boyd and Crawford (2013) best summarize the overall efforts in that it “is less about data that is big than it is about a capacity to search, aggregate, and cross-reference.” Unfortunately, there can be issues that lie in the quality of the data which make it arduous to analyze; this is especially true for social media feeds. Since Facebook and Twitter are personal micro-blogs with unstructured levels of information available, there is no one way to interpret these extensive datasets.

Simply put, users of microblogging are the ‘foot soldiers’ of live reporting. They are the data collectors and suppliers of all that happens around us. Goodchild (2007) has referred to this population as “Citizen Sensors.” While bloggers may not be intentionally sharing their experiences in hopes of providing data for risk assessments, they are nevertheless publicly sharing their activities to socially connect (Stefanidis, Crooks, and Radzikowski 2011). This is particularly true when a post includes a hashtag (the number symbol “#” followed by a key word or phrase that hyperlinks to similar topics). Hashtags, whether topic specific or location specific, allow users to explore what others, even strangers, are saying or doing similarly around them. Therefore, social feeds should be mined and analyzed for improving emergency management
and response. The data is too plentiful to disregard if those producing the information are already openly sharing their experiences; especially, when they include geographical components. Each user’s online testimony can be one less labor intensive step a researcher would need to exert that can delay pertinent information in understanding the mobility of large, potentially at risk, populations. That is why this research explores the benefits social media offers in representing the ever changing real world. As such, a special event model that emulates the LandScan USA dataset is presented at the end with the employment of streamed Twitter posts and Facebook Places check-ins.
CHAPTER II
BACKGROUND

Social media’s locational metadata and content specific information is a captivating quality for population modeling. Its ability to capture activities of population in space and time provides tremendous potential to improve models of population dynamics at high-resolutions. Through spatial and temporal attributes, researchers can begin to understand better ways to approach challenges involved in the medical, political, historical, environmental, social, and technological fields (Kuhn 2012). This class of research alone thrives on instantaneous, spatial information, much of which is provided by everyday citizens. The opportunities to understand how citizens engage in and around their community are endless. From tweeting about opinions of a particular food product, to checking-in to the local gym, a wide variety of individual activity data sets are being currently shared through social media. This section provides a background in the current status and development of geographically focused social media research.

Cyberspace & Neogeography

The evolution from a one-way web browsing experience (e.g. Web 1.0) to now, a two-way interaction and sharing medium is referred to as Web 2.0. Advancements in Web 2.0 platforms have surely transformed the discipline of geography in its most basic concepts of space and place. Some believe this transformation was so revolutionary in its earliest stages that the term ‘Neogeography’ was born to encapsulate a new era of geographic practices through web-based operations (Eisnor 2006, Turner 2006). One of the most influential instruments to Neogeography was the public release of a free, web-based mapping interface, known as GoogleEarth (LeMay 2005). In lieu of expensive cartographic and GIS applications, anyone with access to the Internet could easily upload and share geographic data to make their own maps. This in turn ignited an interest of dynamic mapping to the everyday masses (Turner 2006). Geography was essentially revamped in the public eye with its participation in the growth of Web 2.0. Since then, computer scientists have begun constructing other mapping applications and interfaces to produce maps that are less labor intensive and more automated.
Volunteered Geographic Information

Innovations in mobile Internet access (i.e. smartphones) and authority (i.e. non-experts) are redefining Geography’s main core by blurring the lines of actual space through a virtual space to provide an all access look into the world around us. For example, it has become common practice, particularly in urban areas, to rely on smartphones to browse the Internet for quick information about a place. And, it is also not uncommon for that data we are searching for to be provided by our peers, rather than established experts. This is a result of the Internet morphing into a modernized power structure of bottom up practices, and what Michael Goodchild termed volunteered geographic information, or VGI (2007). Instead of the traditional construction of knowledge, produced and edited strictly by trained professionals, most anyone with an Internet connection now has the ability to share their local knowledge, and in the case of VGI, with locational attributes. This role change of authority allows organizations, businesses, leading entities, and most importantly, non-expert individuals to engage in the distribution of knowledge, ideas, beliefs, and emotions with the simplest post of a tweet (Twitter), status update (Facebook), or picture (both). Geographic information disseminated by citizens ultimately becomes a new avenue of information delivery and consumption. Now, this virtual, two-way street enlightens society by facilitating geographic knowledge in multiple forms by diverse human contributors.

Microblogging, a term born from the fascination of constant posting on social networks, allows the average person to submit information based on their local knowledge of a place and its space. This can be done through various social medias, such as Facebook, Twitter, Foursquare, Instagram, Pinterest, Google+, and YouTube, just to name a few. While some of these entities appeal to individuals in different ways, each proves to be a significant medium for users to learn and inform from their peers or others. This is why it is important to understand the capabilities that social media has to offer in the scientific world. Information is not solely reliant upon experts anymore. Today, it is safe to assume someone has discussed or displayed information about a place that can be useful in some form. It is up to the researcher though to determine just how useful. For this research, analysis covers two of the most popular social platforms, Twitter and Facebook, to understand episodic population dynamics in hopes of modeling average game day fan distributions for emergency preparedness and response.
Twitter

Twitter is certainly one of the industry’s leaders when microblogging is of concern. In fact, as of October 2013, 500 million tweets were sent per day with 100 million daily active users (Twitter IPO filing); this averages to five tweets a day per each individual user. Undoubtedly, during a significant event these figures typically increase. For example, Hughes and Palen (2009) found a 32-58% adaption in ‘active users’ when studying Twitter activity inflations during four significant U.S. events (both Political National Conventions, and two Category 2 hurricanes, Gustav and Ike). This indicates that, during these high profile meetings and extensive natural disasters, a traditional, non-daily user becomes a daily user specifically to participate in informing others of event-related information. Other examples with an underlying increase in Twitter activity, though may not report statistically a difference to the tweeting baseline, are those with a greater concern for crisis analysis (Goolsby 2010; O’Reilly and Milstein 2009; Shklovski et al. 2008) or sentiment analysis (Hoeber et al. 2013, Tumasjan et al. 2010).

Another rationale that demonstrates Twitter as an ideal platform source to harvest media feeds from, is the convenience of its API. On the microblogging service’s website, their policy states that in exchange for using their services, the user agrees to have their information made public and searchable by third parties, as long as the user’s account privacy settings are set accordingly (Twitter Privacy Policy). This agreement is what allows this research to explore Twitter’s 140-character messages to better understand actual, first-hand experiences during a live event. Additionally, the fact that more than three-quarters of active users socially connect through a mobile device only further supports the possibilities of gaining up-to-date access with location accuracy (Twitter IPO filing). However, there is a 1% search limit available through the API’s free streaming service. Plus, social media users are already a sample of the general public since not everyone participates. This digital divide of fans using social media to those who just attend the game day activities and do not use social media, results in these free feeds being an even smaller sample. Nevertheless, any amount of publicly shared experiences or actions, recorded in the moment through smartphones, especially those with locational tracking services activated, can ideally illustrate a more accurate depiction of population dynamics.
Facebook

Facebook is another leader in social media platforms considering its widespread use across personal and business accounts, as well as it being one of the founding social networks that our generation knows. In its origin in 2004 however, Facebook was only accessible to college students from a few selected universities. Now, Facebook connects individuals (13 years of age or older), all across the world with 1.19 billion monthly active users\(^3\). With a user base equivalent to one in every six people on earth, there is no doubt Facebook is a compelling entity for research when it comes to collecting what goes on in the world around us. While Facebook may not be the microblogging platform that Twitter is, it is a more detailed time capsule service with little limitation on what and how people can share their daily activities. Most interestingly, unique venues can be tagged to a user’s post thereby creating a map-blog of where users have been. This geolocation attribute available from Facebook has made check-ins an attractive source of data for researchers.

Unfortunately, such detailed recordings of a user’s activities and location often times prompt stricter privacy concerns, which leads to more challenging harvesting techniques. Krishnamurthy and Wills (2009) discuss how information shared through these platforms can contain some form of personal identifiable information (PII), ultimately compromising the privacy of these users. The first step Facebook has taken to limit these infractions is for Facebook’s Public Feed API to be unlike Twitter’s in that only a few selected media agencies can stream status updates (Public Feed API). There is no free API service available to the public. Additionally, users must have their account profile settings set to “public” in order for shared information to be retrievable; although, this holds true with Twitter as well. These digital barricades are difficult to overstep, which is why the next best spatial and temporal clues to attendee activities are to monitor Facebook Place check-ins. By tracking check-ins to location-specific Facebook Places, a spatial flow of these populations could be interpreted defining movement patterns or local hotspots. Unfortunately, check-outs are not usually provided from

users, therefore these patterns only suggest when and where areas experience increased activity. Nevertheless, knowing where these fans move in and around the football stadium will further support the accuracy of emergency planning. And when time is a key factor in emergency response, check-ins through Facebook can suggest which areas are of most interests.

**Limitations of Social Media**

It should be understood that the information being provided on social media are not from traditional subject matter experts, but rather, from the general population. Because the data could be harvested in real-time for some applications, there would be no time for expert knowledge to demonstrate the distributions of episodic populations at special events without social media. As such, it is crucial to explore the benefits of incorporating social UGC. In this study, those providing the information are most likely football fans, not emergency responding experts or research professionals. They might not even be aware that their basic tweet or check-in gives insight into quantifiable data of how people move to and from the football stadium. Their motivation to share their current endeavors is not necessarily persuaded by a responsibility as in VGI, but rather to display their lifestyle (Bhattacharjee and Miah 2011). And thus, data may not always be coherent or spatially accurate. As non-professionals, these users are providing large amounts of data that are “inherently noisy” (Becker, Naaman, and Gravano 2010) which requires supplementary efforts to interpret.

In order to formulate a consistent dataset for analysis, local knowledge was administered to help filter through this sample of unstructured Big Data. The benefit of the preceding statement was to eliminate any material that is certainly wrong. According to Elwood, Goodchild, and Sui (2012), because a large amount of unstructured data is generated, it also “presents a number of challenges for developing methodologies.” Without an authoritative structure to restrict erroneous observations, information can produce false-positive results skewing representations of population dynamics. Tweets are notorious for this in their very nature due to such limited posts that can be incorrectly interpreted or, more importantly, untruthful. It was assumed in this research that the intentions of users were to honestly share information about their currently activities, at the appropriate moment in time. For that reason, social media was explored here, as an additional variable in the methodology of producing future
models of episodic populations. The study was merely a near, real-time analysis of game day activities analyzed throughout the season and verified against other reports and statistics. Or in other words, it was an assessment of the feasibility of social media as another input source for population modeling at high-resolutions.

Applications of Social Media Data

Due to the many concerns of privacy on public sharing services, there are very few opportunities to have access to complete or comprehensive data streams. There have been many studies done that model who the users are, what their characteristics are, who they connect with, what they do, and where they reside. Depending on the information retrieved, researchers can model different aspects of the world around us. Some of the best work to date, investigate the influence social media has on the general public. For example, Sakaki, Okazaki, and Matsuo (2010) dubbed Twitter users as “social sensors” that offer a unique look into where affects were felt following earthquakes that occur in Japan. Those updates provided insight into where the true disaster zone actually was. These tweets can also provide faster information about an earthquake than the official Japan Meteorological Agency (JMA) can, being that the JMA has authoritative procedures to follow first before a statement can be issued. Social media can also alter the normal patterns of life’s activities making it necessary to understand a new or current way of life. When the Tunisian Revolution unfolded in early 2011, the documentation by its citizens on Facebook of the chaos that was igniting in the streets, ultimately, led to an invitation made through Facebook to plan a protest (Marzouki et al. 2012).

Another realm of this research, not necessarily centered around exact spatio-temporal modeling, examines just the semantics of a user’s post to understand what users are feeling or supporting and where. By creating linguistic parameters, researchers narrow in on where individuals are by searching for locational clues or by keywords related to a trending topic. In 2010, Cheng, Caverless, and Lee analyzed Twitter content to estimate a user’s location at city level by identifying words that tend to have a geographical scope (i.e. “howdy,” a Texas greeting word). Other studies only skim the surface and locate users at city-level based on their friendship connections (Rout and Preotiuc-Pietro 2013). This approach proves too vague or time consuming because of its qualitative design and is better suited for interpreting public emotions and opinions.
Based on a post’s sentiments. Examples of these studies include: trying to predict outcomes of national events, such as presidential elections (Schoen et al. 2013), Bollen, Pepe, and Mao’s study in 2009 of overall national moods during major events (i.e. economic depression), and the potential well-being of others or communities (Kautz 2013).

One example of this stagnant modeling process is a special event model of San Francisco surrounding AT&T field where the San Francisco Giants play. The area of interest was the upper east end corner of the city’s peninsula, where the PG&E gas pipe lines are situated. An assessment was needed for a worst case scenario of the possible at-risk populations within this area. At the time of the request, the 2012 World Series in major league baseball was in effect. Therefore, the analysis incorporated the excessive number of fans participating in the event’s activities to the baseline population of the city’s worker population. Information was gathered, retrospectively, through news articles, transportation reports, blogs, stadium statistics, areal images and personal pictures, as well as pinpointing the locations of local gathering spots such as restaurants, bars, and parking lots. From there, populations were distributed throughout the research area two hours prior to the first pitch, and during the middle of the game around the 4th inning when population counts were assumed to be at a maximum (Figure 2). This workflow of data gathering, while accurately verified, is too time intensive for this study and requires research over many sources. Additionally, some of the sources are again, UGC and provide the same information that social media can on the fly. The importance of incorporating tweets and check-ins to population modeling dynamics is what makes this retrospective approach a faster process. As a result, high-resolution population modeling of episodic populations can be implemented faster, thus speeding up the process of emergency management and response.

**Nature of the Problem**

There are many ways to model users by interpreting their social patterns, but there has still been no effort to incorporate these unauthoritative approaches to modeling at risk populations in near real-time. While previous studies in disparate disciplines have addressed that there are explanatory capabilities of existing social media data, few have specifically addressed how it could be utilized for models depicting population distributions and dynamics. Though there is much noise in this type of unstructured data, it often contains detailed space-time information
Figure 2. An example of ex post facto modeling techniques of a San Francisco Special Event Model 2 hours before game (left) and during game (right) when attendance is assumed at its maximum.

unavailable from any other source, which is crucial when it comes to understanding near real-time population dynamics at high-resolutions.

Considering the example of the special event population data development from San Francisco discussed previously, it is clear that developing such population distribution scenarios with data from past events may not always provide necessary information for an upcoming event. More importantly, for emergency management, total population at a special event location, for example a stadium, is not adequate for optimizing event management measures. Information on when and where large groups of population congregate around the event venue and possible patterns of flow into the venue are extremely valuable for understanding population distribution and dynamics for a special event and useful for emergency response planning. Such near real-time information could only be gathered from widespread in situ sensors (such as public cameras) or from harvesting social media platforms. While many data generation techniques from social media exist, no one has attempted to incorporate social media data as an input to high-resolution population distribution and dynamics models.

Consequently, this research assesses this feasibility of integrating social media feeds. In order to understand the mobility of these attendees, attributes that define a home game for the 2013 University of Tennessee Football season, plus Twitter posts and Facebook Places check-ins are compared. With local knowledge of the university’s football program and game day activities
and, a one and a half mile study area around the football stadium, assessments are made from the social activity 24 hours surrounding each home game. Through this streamed data, mobile patterns and hot spots arise that prove popular for tailgating venues. These indicators suggest when attendees begin showing up to the university’s campus, when the game begins, when significant events occur during the game, and when the game is over and how the crowds travel home. Cues such as these are what can set social media influenced models apart from the time-extensive current approaches.

**Objective**

Given the limitations of assessing special event population dynamics discussed previously, the overall objective of this research is to address:

- *The feasibility of integrating social media as an input data source for a population distribution and dynamics model of special events.*

By taking an exploratory approach to observations from social media, better representation of population dynamics may be possible for populations that are normally not accounted for in a population distribution model. Information on space (location) and time (temporal) contained in social media data could be exploited to assess the characteristics of episodic populations. In order to test these assumptions, this research particularly focused on Twitter’s tweets and Facebook’s check-in data collected during the 2013-2014 football season at the University of Tennessee (UT) with the following sub-objectives:

- *What is the relation between social media data to total reported attendance at the stadium?*

- *Are there particular spatiotemporal patterns in the data that provide insights to refinements of the LandScan USA Special Event population distribution model?*

The first statement is easily obtainable because these games are highly publicized and recorded, and can therefore be easily verified to reports that reference statistics during the game. Plus, the multiple occurrences of these games allows for comparisons between each game and those verified sources. However, while these games are seasonal, weekly occurrences, they are also singular events which can reveal any unknown population movements that usually occur with episodic populations. For instance, these games do not always occur at the same time of the day
or take place each consecutive weekend. They may not experience the same type of weather, and they do not even appeal to the same crowd if the football team is doing poorly or is playing a particular opponent. These many unshared features are what make exploring a reoccurring event, such as UT’s football games, essential for assessing the feasibility of social media. They are unique instances in a series of one particular event, which can offer different trends or anomalies that demand near, real-time updates from social media to validate.
CHAPTER III
METHODS

To assess and demonstrate how social media can enhance population distributions for special events, four main steps were implemented, and each is summarized here. First, data collection was performed by two other colleagues to harvest Twitter’s tweets and Facebook’s check-ins. The next step was a manual collection of ex post facto reports about the game day’s characteristics to evaluate against the social media data; some of which were also compared to local knowledge. Following that was the process of organizing the combined datasets to create a standardization of the data that would be quantifiable to establish episodic signatures and to process spatial analysis. And finally, an output raster was created displaying the average population distributions on game days, based on the results from this research.

Streaming Collection

Tweets and check-ins were gathered through two forms of API streaming in a predefined time and locational buffer. In both platforms, the entire game day’s feed was captured from 12 hours before to 12 hours following each game’s kickoff. For Twitter, data was received compiled together in total counts of 15 minute increments in order to simplify the enormous amount of tweets returned (see Appendix A.1). The complete 24 hours produced a timeline of events with spikes and stagnation that could represent a moment of significance worth comparing to factual documentation (see APPENDIX B). Facebook on the other hand, was solely an extraction of the latest number of check-ins to a Facebook Place’s page every 30 minutes (with the exception of Neyland Stadium which refreshed every 15 minutes) (see Appendix A.2). Another timeline was produced, but this time representing when people moved about, and is displayed by the hourly percent change in check-ins for each game (see APPENDIX C). These fluctuations in the temporal activity were important for understanding which times were most relevant or when popular activities might be occurring, so that proper population dynamics could be understood.

The study area for this research examined many scales to perform global and local spatial analyses. For most of this paper, a buffer of 1.5 miles surrounding Neyland Stadium remained the main focus in order to create a special event distribution model at high-resolution, though
multiple, increasingly larger buffers were also applied. For Facebook check-ins, 95 Facebook Places were identified within this 1.5 mile buffer as areas of interest to game day fans (e.g. restaurants/bars, tailgating locations) (for full list, see APPENDIX D). Each was monitored throughout the day, as stated previously, in 30 minute increments to create a timeline of checkpoints for moving populations. Additional Facebook Places were located within this buffer (e.g. Hospitals, law firms, and art museums), but were not selected based on their irrelevancy to game day fans.

Twitter’s mining parameters, however, were much larger, capturing data for the entire world in order to gather tweets that were not necessarily geolocated. To determine the appropriate tweets, a compilation of 35 key words associated with the University of Tennessee or its football team (e.g. Vols, Team117, GBO), plus additional weekly opponent terms were identified to filter the returned tweets (for full list, see APPENDIX E). Through Twitter’s API, tweets were collected from the feed’s free, firehose sample stream of 1%, per request. At an additional cost, the entire stream of tweets is available, but was not considered for this research. Because Twitter’s feed returned many unique, 140-character messages, geolocated tweets were then categorized into multiple scales of 1.5, 5, 10, 20, 50, and 510 miles. The smallest radius of 1.5 miles, used in both platforms, provided a finer spatial resolution of fans attending the game day’s activities by encompassing all of the university’s campus, as well as local surrounding areas of interest. The other distances included geographical attributes, such as: 5 mi. for adjacent neighborhoods and businesses, 10 mi. for almost all of the City of Knoxville, 20 mi. for all of Knox county, 50 mi. for all of the surrounding eight counties, and lastly at 510 mi. for all of the opponents’ home stadiums (with the exception of the University of Oregon). Preliminary analysis was performed with the worldwide coverage of tweets to understand other emerging global patterns, but offered no additional benefits to modeling special event populations and were eventually ignored.

Collectively, these two platforms documented how game day dynamics fluctuate around the stadium and campus. Movement signatures were produced for each game in the smallest study area of 1.5 miles using check-ins and geolocated tweets. Twitter’s raw data of tweets, that are not necessarily geocoded, produced activity signatures for each game that represented when moments of significance occurred during the 24 hour period. These types of signatures averaged
over the season, are what ultimately help in the definition of refining a population model to a special event high resolution distribution dataset. However, these data do not include the entire season’s games due to unanticipated computational reasons. What was captured, was still sufficient and is explained next.

Collected Data

The 2013-2014 University of Tennessee football season included a total of twelve games, consisting of seven home games, and five away games (Figure 3). For purposes of this research, Twitter data for only eight of these scheduled games – four home and four away – were analyzed. Unexpected server restarts for system updates caused the data collection software to stop unexpectedly for the remaining four games. Conversely, Facebook check-in data, for all twelve games, was successfully analyzed from 95 different Facebook Places surrounding the home stadium. However, no Facebook check-in data were obtained for the final three games. This could have been caused by a malfunction of the automated data collection program or irregularities of Facebook servers; neither of which was further investigated. Thus, in order to minimize the effect of the aforementioned inconsistencies, which would inevitably skew the results of this study, (1) each social media data were analyzed separately, and (2) only that data which was collected consistently and comprehensively was considered. This section discusses the results of those analyses, as well as provides a discussion about how social media could assist in the analytics of population dynamics for special events.

![Attendance Totals for the 2013-2014 University of Tennessee Football Schedule](image)

Figure 3. 2013 Season outlook for the University of Tennessee’s football game schedule.
From the first game, held on August 31, 2013, to the last game, held on November 30, 2013, a total of 2,076,875 tweets and 904 check-in data were collected. These were captured in a twenty-four hour time period, spanning 12 hours before and 12 hours after each game’s kickoff. Geo-tagged tweets accounted for only 2.8% of the total tweets collected from Twitter, as opposed to Facebook’s check-ins which were 100% geolocated since each Facebook check-in is by design associated with an actual venue or geographic place. This percentage of geo-tagged tweets falls within the percentage of tweets with GPS coordinates found in other studies, which ranged from 0.47% (Cheng 2010) to 3.17% data (Morstatter et al 2013). These aforementioned reports are for tweets that were specifically geo-tagged, and not based on if a user’s profile description provided a “Home” location. The purpose for this was to know where the actual tweet was being produced and not where the user resides permanently.

Out of the 35 UT specific terms used in the search criteria, seven particular terms—which will be referred to as the “Top 7” for the remainder of this paper– comprised over two-thirds of all the returned tweets. These terms consisted of the top six most used terms, plus “Neyland” (the 16th most used term), which was essentially useful for its locational significance to the home stadium. This refinement of tweets still produced 1,408,139 total tweets over the eight games (home and away), wherein 22,748 (1.62%) were geolocated. Though the Top 7 dropped the number of geolocated tweets, the percentage still remained within the range of other past studies and was therefore assumed to be sufficient for this research. As such, to model population at high-resolutions, analysis for this research was narrowed to include only the following seven terms: “Football Time In Tennessee,” “Tennessee,” “Vols,” “GBO,” “VFL,” “Big Orange,” “Neyland.”

**Verification & Validation**

In order to support and verify the collected social media data, it was essential to collect supplementary open source data to compare game day population dynamics and significant moments during the game. Newspaper articles, broadcast reports, pictures, and traffic
alerts/messages were all explored, but the most useful source derived from the university’s athletic department’s website\(^4\). Under the football team’s section of the website were individual game day statistics, as well as a seasonal review of that year’s schedule and total attendances. These data were valuable to establish a timetable of certainty through factual documentation and summaries of the day’s characteristics. Furthermore, a documented summary of scores produced throughout the game, team and individual player’s statistics, as well as the game’s start time and duration were reported in what is known as the ‘Box Score’\(^5\). This temporal information confirmed much of Twitter’s activity, as discussed previously (again, see APPENDIX A), by supporting the times of increased number of tweets during moments such as kickoff or a team scoring. It also provided a guideline for asserting when population hot spots or assembly points should be measured before or after the game in session.

An additional avenue of supporting data pursued was through the school’s athletic department to receive the spatiotemporal data produced by ticket scanners positioned at each entrance gate of Neyland Stadium. For each game, as many as 100,000 football enthusiasts must have their tickets scanned by a digital ticket reader in order to enter the stadium. This process creates a precise timestamp, within two hours prior to kickoff when the stadium’s gates open, of where and most importantly, when fans have moved from the surrounding areas to inside the stadium. Initial hope was to receive these reports to validate as ground truth to Facebook’s check-ins of Neyland Stadium (collected every 15 minutes), however, this information was not made available to the public and therefore was not possible to use for this research. This information, reflecting transactions of scanned tickets for stadium entry, would have improved the distribution of this model. Instead, confidence from the university’s Box Score and attendance reports were the only additional sources for validating game day dynamics.

In addition to the unobtainable data from ticket readers, or actual access to complete data feeds, missteps in the collection process and lack of data retrieved occurred. These setbacks were not planned, but possibly could have been avoided. From unrefreshed search criteria to server

\(^4\) http://www.utsports.com/sports/m-footbl/sched/tenn-m-footbl-sched.html
\(^5\) http://www.utsports.com/sports/m-footbl/stats/2014-2015/ut...(0831)...html#GAME.NEW
shutdowns, data was not entirely collected. These errors hindered the research to its fullest capacity, and should be avoided if this research is to be repeated.

**Limitations of Twitter**

Unfortunately, data collection was inconsistent for Twitter. For instance, in four (games 5,6,11,12) of the 12 games, the server that collected from Twitter’s API failed, thus, no data were collected. One workaround was to retrieve those games from a 1% subsample of the data collected on another server. However, this caused a drop in the data that were returned since the 1% was an open ended search criteria for all of Twitter. Additionally, the search criteria for opponent’s term for games 3, 9, and 10 were not refreshed on and instead was residual from the previous week’s opponent’s terms (again, see APPENDIX A).  

Another unforeseen outcome was the false-positive that came from the keyword “UTK.” This acronym, which stands for the ‘University of Tennessee, Knoxville,’ is also a real word in the Indonesian language (Bhasa Indonesia) meaning “for”. This was noticed after the second game when many of the geocoded tweets were locating mostly in Indonesia (Figure 4). This false-positive data presumably skewed overall results, for example, as the amount of positive “UTK” tweets accounted for 36% of all the returned tweets (27% geocoded tweets) during the first game.

While this false-positive keyword had the greatest impact on the data, it was not the only one. Other’s included the university’s sports teams being referred to as “Volunteer(s),” and their mascot’s name being “Smokey.” These two terms were used throughout the world because of their wide spread use in many different contexts. Other examples of terms that might be used in several contexts include “orange,” “rocky,” “pregame,” and “kickoff.” Nevertheless, these terms remained in the overall search criteria, and only “UTK” was modified to include a hashtag (i.e. #UTK) to avoid misuse of the term. As a result of the implementation of this particular hashtag,

---

6 For games 3 and 9, fewer tweets were returned because of a non-nationally recognized game (WKU during week 3), and a BYE week (Alabama did not play week 9). More tweets were returned for the Auburn game (10) because the opponent’s search criteria for game 8 (Alabama) was still in effect; the increase in tweets was a result from Alabama playing a historical rival, and highly ranked, LSU.
there was a significant drop in returned tweets associated with “#UTK.” This drop in returned tweets indicated that the term was not, in fact, as popular of a keyword in conjunction with the university’s football games as originally identified.

**Limitations of Facebook Check-in Data**

Out of the 95 available Facebook Places, check-ins were restricted to less than half of those places. This is surprising considering each of these places were within a 1.5 mile radius of the stadium. Neyland Stadium’s Facebook page was otherwise very active. Figure 5 represents how spectators begin to show up to the stadium 2 hours before kickoff, when the gates open to the public, to half an hour following kickoff. Even though there were reported check-ins to the stadium before gates opened, these were discarded to adjust for actual entries to the stadium. The 5 home games that reported check-ins, were normalized to each game’s maximum check-ins so that each game was comparable to the movement dynamics of other games.

This figure supports the notion that despite when the game starts, or whether the opponent is an important rival or unranked team, check-ins to the stadium are mostly consistent. Movement is slow in the first 30 minutes the gates are open, as well as the last 30 minutes
following kickoff. Most of the influx occurs within an hour before kickoff. There is some lag time between the earlier season games and the mid-season games, as to when people start heading into the stadium (e.g. roughly one hour before kickoff), but otherwise each game shares a similar signature for arrivals. These results are a defining indicator that the university’s fan base adopts a reliable trend into the stadium which can support predictive trends at a small scale. The challenge is increasing the scale to support areas outside the stadium that can produce consistent movement trends. Because Facebook check-ins were too few elsewhere, geolocated tweets are the next, best option, but will require additional spatial analysis to confirm their relevance to the university’s campus.

**Standardization & Testing**

With the bottom up approach to microblogging, information is not only unique to individual users but, also inconsistent. This presents a time-consuming obstacle for researchers to analyze,
specifically with Twitter because of the magnitude of cryptic and very specialized information being provided (Hoeber 2013). Therefore, the metadata (i.e. timestamp, lat/long) was investigated to accurately understand the patterns and clustering of tweets, and less about the sentiment. This was imperative considering the other source data relied on check-in counts, which only monitored a specified channel for increasing number of actions (e.g. the number of people checking in to a place). With the main focus on geolocated tweets and check-in counts, exact locations were defined in particular areas of interest. The greatest emphasis is directly surrounding the stadium in the 1.5 mile buffer since this area is the most important when modeling this special event attendance. However, because geolocated tweets were collected outside of this 1.5 mile buffer, two spatial analyses were performed to statistically depict population dynamics of the game day’s fans.

The first statistic is the Moran’s I test (Moran 1950). Moran (1950) first proposed this statistic to test the degree of spatial autocorrelation between two observed locations with a specific set of features. Spatial autocorrelation characterizes how variables relate to one another as “more similar (positive correlation) or less similar (negative correlation) than expected for randomly associated pairs of observations” (Legendre 1993). Therefore, the Moran’s I statistic incorporates added attributes to test against the null hypothesis that the data are independent and of random process (Li, Calder, and Cressie 2007). The alternative hypothesis, which was the goal for this research, states that the data are clustered and not of random chance. Using the Moran’s I method implemented in ArcMap’s Spatial Autocorrelation tool, geolocated tweets were summed together to a cell size of 3 arc seconds (mirroring that of LandScan USA) to create an observed location with features, then analyzed as a whole at multiple scales as discussed above. The entire dataset is analyzed because these statistics are global in that they require measurements from all of the features to assess the overall spatial autocorrelation (Ord and Getis 1995).

The second test is the Local Indicator of Spatial Association (LISA) (Anselin 1995). This statistic differs from the Moran’s I because it focuses on local associations of each observed variable and “allows for the decomposition of global indicators” to define clusters and outliers that are statistically significant (Anselin 1995). It is important to follow up with this test to interpret where, and not so much how, the geolocated tweets relate. Using the LISA method implemented in ArcMap’s Cluster and Outlier Analysis tool, tweets are again summed to a 3 arc
second cell and analyzed from one cell to neighboring cells. A raster output is given indicating those cells which are hot spots, cold spots, or outliers. The goal in this research was to find those areas which are hotspots so that they may be used as a basis for modeling game day fans at high resolutions. The tool was performed at multiple scales, but was only most necessary at the 1.5 mile buffer for creating a special event population dynamics model surrounding Neyland Stadium (explained in the next section).

**Model for Future Use**

The final step of this research produced a prototype of LandScan USA as a Special Event population distribution layer for the University of Tennessee during two times of the day: a) 2.5 hours before kickoff, b) 30 minutes after kickoff. Though this model imitates the original LandScan USA model, there will not be an average expected daytime or nighttime distribution. Instead, these two distributions will represent the typical dynamics before Neyland Stadium’s gates open to the public and right after kickoff when it is assumed that all ticketed fans are in their seats. The purpose of this is due to the fact that these games occur at different times which can occur during the day or at night. Therefore, this model is simply an extension for the downtown Knoxville area during a special event; which in this case, is an average football game day (Saturday) in the Fall. It spatially portrays a specific expected representation of where football fans gather, while also identifying areas that are of most significance for game day activities. To understand how the areas around the campus increase in population, the Special Event model is eventually compared to the LandScan USA nighttime baseline.

The inputs for creating this model will derive from the results of this research. Tweets and check-ins will weigh the areas around campus that represent observed fan distributions. Additionally, distance will be a factor in the input methods for placing stronger weights around campus and closest to Neyland Stadium. Two models will be produced depicting two different times during a football game day, and therefore will require two beginning special event population counts that should represent the respective times during the day. Each count will comprise of fans attending the games, as well as fans not attending the game. Those not attending will consist of fans without tickets to the game, worker populations for the surrounding
businesses, law enforcement officials maintaining game day activities, and residential populations.

Though this model mimics the LandScan USA model, other population datasets and methods could have been used. This parallel was chosen considering LandScan USA is the most up to date and finest resolution dataset for this area. It also allows for the most realistic baseline population since the nighttime layer is likely the most equivalent distribution on the weekends. And while the goal of this research is to assess the capabilities of using social media to aid in the creation of special event distribution models, the long-term intention is that this prototype be the avenue to which other models can be implemented on-the-fly for other special events across the nation. Understanding the mobility patterns of these event attendees is necessary since they are not normally accounted for and can be the most affected during emergency situations. Local emergency response teams would therefore benefit immensely from this type of localized information and can potentially offset lurking threats. Furthermore, interested parties like the university’s marketing sponsors, restaurant and business planning committees, or even traffic operations and support can benefit by knowing how and where certain areas increase in population.
CHAPTER IV
RESULTS & DISCUSSION

The following sections describe the spatial and temporal trends observed in the Twitter and Facebook data, as well as a discussion of their benefits and limitations in the research. Prominent features in these trends (peaks and troughs) were correlated to reported event’s data, such as a game’s kickoff time or the stadium’s attendance. In addition analysis of the geolocated data using the Global and Local Moran’s I tests were performed to explain any spatial clustering of tweets and how reliable they are to the episodic fans of the university’s football games. Finally, a LandScan USA equivalent raster (represented in 3 arc second or ~90m cells) is presented depicting two times during a football game day using the results and interpretations of each test.

Tweets and Check-ins

Social media data from home games were analyzed to explore possible correlation between social media data and stadium attendances. A non-zero correlation between the two would allow for some predictable measures of social media so that if it were not fully captured in future studies, estimated trends can be inferred; as for example, what happened with four of the 12 games this year. However, it is important to note that the data received from social media comes from a fraction of the population around the stadium since not all fans are actively involved with social media. Nevertheless, the data are valuable because they are one of the few forms of open data streams that bear clues to location and possible behavior of population around a place. Such data can also be analyzed against official university reports of stadium attendances and local knowledge.

The attendance counts used for Neyland Stadium are derived from the university’s Scoring Summary report\(^7\) found on the football team’s website. Attendances are reported from ticket sales, rather than the digital ticket readers scanning each individual into the stadium; as stated earlier, that information was not available to the public. Consequently, there is most likely

some error (overestimation) in the difference of tickets sold to actual fans attending the game. Because true attendance data were not available, it was assumed that ‘tickets sold’ and attendance at a game were the same. Additionally, there is most likely some underestimation as stadium workers and team affiliated staff members are not counted in the attendance report.

The following two figures represent the total number of tweets and check-ins (the dependent variables) captured in their respected games, in relation to the attendance at each home game (the independent variable). Figure 6 is the total number of tweets using the Top 7 terms for four of the seven home games, reflecting a total of 738,541 tweets. Those games are Games 1, 2, 7, and 9.

![Home Games' Tweet Totals to Stadium Attendance](image)

**Figure 6. Total number of tweets for each home game compared to that respected home game’s attendance.**

Figure 7 is the total number of check-ins for the first five home games (Games 1, 2, 5, 6, and 7), as data from the last two home games, as stated previously, were discarded due to the unexplained, and likely erroneous, total of zero check-ins.
Both the numbers of Tweets and Facebook check-in have too few observations (games) to derive a truly conclusive statistical correlation. Rather these provide overviews of social media signatures. Figure 6, shows, at a minimum, there are around 150,000 tweets produced each game in a 24 hour window. These numbers are substantial in providing many insights about game day dynamics, and suggest that each game could produce large quantities of social media and information contained in them. Figure 7 might suggest (excluding the two zeros) a potentially positive correlation with check-ins increasing as attendance increases, but still requires additional observations to conclude there is predictability in this data. Attendance at a game, population dynamics around a game venue, and social media generation depend on many different factors such as weather, seasonal performance of a team, attractiveness of an opposing team etc. With data collected consistently and comprehensively over a long period of time (such as several football seasons in this case) that reasonably accounts for normal variability on those factors, the true predictive power of social media data for modeling episodic (special event) populations can be evaluated.
Facebook Results

The benefits from Facebook check-ins allow for an understanding of when and where fans begin to show up around the campus to participate in game day activities. This type of information is crucial when modeling population dynamics for special events because it creates a window of when the study area has an influx of not normally accounted for populations. Figure 8 (as well as the figures in APPENDIX C) represents this movement signature by depicting each game’s check-in activity around Neyland Stadium. Most of the initial increased activity occurs within 6 hours of kickoff. Then, check-ins begin to stall right after kickoff. This stagnation would suggest that most of the special event population would be in a designated area for the duration of the game and not moving around campus. Following the game, however, the check-in activity begins to increase once more three to four hours after kickoff suggesting that the game has ended and fans are moving about again. This signature increases, generally, up to six to eight hours after kickoff, then begins to stall thereafter, with the exception of the first game versus APSU; the only explanation for that game is possibly due to the fact that it was the first game of the season, as well as the win for UT, and excitement must have been high even in the early hours of the following day since this game started at 6:00pm.

![Total Facebook Check-ins for Each Game](image)

**Figure 8.** A seasonal outlook of each game day’s check-in activity 12 hours before and 12 hours after kickoff.
Another note worth mentioning, is the fact that check-in activity began to drop in popularity as the season progressed. The first game had a total of 330 check-ins, while the last game with check-in activity only produced 15 total check-ins; potentially, Facebook check-ins might have been obsolete altogether as the last two games reported zero check-ins. The only game that opposed this seasonal decline was the sixth game of the season, versus UGA. That week’s game was popular enough to total 161 check-ins; the second most active check-in game of the season. This may be due to the opponent being a ranked team that UT almost won against, but the check-ins before kickoff would suggest that is was already a popular game day and many fans were already excited even before a potential win.

One unforeseen outcome with Facebook check-in data, was the lack of Facebook Places actually being checked into during game days. Out of the 95 Facebook Places identified, not even half were used once. The most venues used during the studied 24 hours occurred on the first game and only accounted for 43 different Places (including Neyland Stadium). Throughout the remainder of the season, the number of venues for check-ins dropped to around the mid to high 30s, then to just 22 for Game 7 (see figure descriptions in APPENDIX C). With declining check-in totals and venue usage for each passing game, Facebook data was only useful for most of the season. The lack of popularity may have been the result to a similarly declining winning percentage for the football team, but was not explored further since that hypothesis may only be answered through semantic research of social media.

**Popularity of Tweeted Terms**

Tables 1(a-c) display the popularity of each term, 12 hours leading up to kickoff and 12 hours following kickoff, in one hour intervals for: all of the 8 games (1a), all the home games (1b), and for all the away games (1c). Values for individual terms were normalized by dividing the hourly values by that term’s highest count in the 24 hour period. Additionally, each hour interval is represented by one of five colors depicting the term’s usage percentage respective to that day’s maximum usage. The lighter oranges show less usage, while the darker oranges show the most usage.

Normalizing is necessary to distinguish each term’s signature throughout the day for different game scenarios, and to compare its usage amongst the other terms. For example,
Table 1. (a,b,c). Twitter’s Heat Map of the ‘Top 7’ terms in an aggregate 24 hour window for multiple game type scenarios.

<table>
<thead>
<tr>
<th></th>
<th>Legend</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0 - 0.199</td>
</tr>
</tbody>
</table>

### a. All games (8)

| Term                                | -12 | -11 | -10 | -9 | -8 | -7 | -6 | -5 | -4 | -3 | -2 | -1 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|-------------------------------------|-----|-----|-----|----|----|----|----|----|----|----|----|----|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Football Time In Tennessee          |     |     |     |    |    |    |    |    |    |    |    |    |   |   |   |   |   |   |   |   |   |   |   |   |   |
| Tennessee                           |     |     |     |    |    |    |    |    |    |    |    |    |   |   |   |   |   |   |   |   |   |   |   |   |   |
| Vols                                |     |     |     |    |    |    |    |    |    |    |    |    |   |   |   |   |   |   |   |   |   |   |   |   |   |
| GBO                                 |     |     |     |    |    |    |    |    |    |    |    |    |   |   |   |   |   |   |   |   |   |   |   |   |   |
| VFL                                 |     |     |     |    |    |    |    |    |    |    |    |    |   |   |   |   |   |   |   |   |   |   |   |   |   |
| Big Orange                          |     |     |     |    |    |    |    |    |    |    |    |    |   |   |   |   |   |   |   |   |   |   |   |   |   |
| Neyland                             |     |     |     |    |    |    |    |    |    |    |    |    |   |   |   |   |   |   |   |   |   |   |   |   |   |

### b. Home games (4)

| Term                                | -12 | -11 | -10 | -9 | -8 | -7 | -6 | -5 | -4 | -3 | -2 | -1 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|-------------------------------------|-----|-----|-----|----|----|----|----|----|----|----|----|----|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Football Time In Tennessee          |     |     |     |    |    |    |    |    |    |    |    |    |   |   |   |   |   |   |   |   |   |   |   |   |   |
| Tennessee                           |     |     |     |    |    |    |    |    |    |    |    |    |   |   |   |   |   |   |   |   |   |   |   |   |   |
| Vols                                |     |     |     |    |    |    |    |    |    |    |    |    |   |   |   |   |   |   |   |   |   |   |   |   |   |
| GBO                                 |     |     |     |    |    |    |    |    |    |    |    |    |   |   |   |   |   |   |   |   |   |   |   |   |   |
| VFL                                 |     |     |     |    |    |    |    |    |    |    |    |    |   |   |   |   |   |   |   |   |   |   |   |   |   |
| Big Orange                          |     |     |     |    |    |    |    |    |    |    |    |    |   |   |   |   |   |   |   |   |   |   |   |   |   |
| Neyland                             |     |     |     |    |    |    |    |    |    |    |    |    |   |   |   |   |   |   |   |   |   |   |   |   |   |

### c. Away games (4)

| Term                                | -12 | -11 | -10 | -9 | -8 | -7 | -6 | -5 | -4 | -3 | -2 | -1 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|-------------------------------------|-----|-----|-----|----|----|----|----|----|----|----|----|----|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Football Time In Tennessee          |     |     |     |    |    |    |    |    |    |    |    |    |   |   |   |   |   |   |   |   |   |   |   |   |   |
| Tennessee                           |     |     |     |    |    |    |    |    |    |    |    |    |   |   |   |   |   |   |   |   |   |   |   |   |   |
| Vols                                |     |     |     |    |    |    |    |    |    |    |    |    |   |   |   |   |   |   |   |   |   |   |   |   |   |
| GBO                                 |     |     |     |    |    |    |    |    |    |    |    |    |   |   |   |   |   |   |   |   |   |   |   |   |   |
| VFL                                 |     |     |     |    |    |    |    |    |    |    |    |    |   |   |   |   |   |   |   |   |   |   |   |   |   |
| Big Orange                          |     |     |     |    |    |    |    |    |    |    |    |    |   |   |   |   |   |   |   |   |   |   |   |   |   |
| Neyland                             |     |     |     |    |    |    |    |    |    |    |    |    |   |   |   |   |   |   |   |   |   |   |   |   |   |

“Football Time In Tennessee” is a phrase often used at the moment of kickoff, as seen in 1a and 1c. The phrase is announced over the intercom at Neyland Stadium as the kicker kicks the football in the opening play. It was assumed that this would suggest when the phrase was used most often; however, during home games (1b), this was not found to be consistent. Instead, the peak usage is in the fourth hour interval following the games’ kickoff. This may suggest that the phrase is a positive expression used when games are won (e.g. three out of those four were wins), or it may suggest that the games were early in the day and fans continued to use the phrase (e.g. three games started at twelve o’clock in the afternoon). It is also possible that both of these hypotheses are collectively responsible for the increased during the fourth hourly interval after kickoff.

In contrast, away games (1c) have less breadth of Twitter activity compared to home games (1b). The fact that the results from all games combined (1a) resemble the home game patterns, as compared to the away games, probably indicates that the intensities of Tweets during...
home games were significantly higher than away games. By creating these signatures, a more refined focus can be implemented when understanding the activeness of episodic populations, and less time can be wasted modeling hours that are underrepresented. The signatures may also imply which times are most important, in addition to the duration of time necessary to obtain the best data. Instead of studying a 24 hour window surrounding kickoff, maybe a 12 hour window would have allowed for more efficient results.

**Spatial Significance**

The global and local clustering of geocoded tweets based on the Top 7 terms were statistically analyzed using ESRI ArcMap’s Global Moran’s I tool and the Local Indicators of Spatial Association (LISA) tool. Geocoded tweets were analyzed for spatial autocorrelation and significance using multiple radiiuses of 1.5, 5, 10, 20, 50, and 510 miles from Neyland Stadium. The purpose of using these tests at different scales was to assess the usefulness of the search criteria that was performed by determining if patterns exist, and if so, which areas were hotspots or outliers. In order to spatially represent the university’s fan base relevance, it is important to understand these geographic measurements. Appropriately, these spatial statistics can provide potentially predictive population dynamics indicators. These exploratory approaches of analyzing social media to predict episodic population hotspots will be beneficial for future research as more information is available through these platforms. For now, the results from the LISA test, and the locations of the Facebook Places, are applied to create a raster output for the university’s game day population dynamics model.

**Global Moran’s I**

The first test was performed using ArcMap’s Spatial Autocorrelation tool. This uses the Global Moran’s I test which measures the spatial autocorrelation between a feature’s location and attribute occurrence or value, at the same time. For the purpose of creating a raster output similar to that of ORNL’s LandScan USA distribution data, tweets were summed together to a gridded cell size of 3 arc seconds, then snapped to the LandScan USA environment. Therefore, this test assessed whether these cells with tweets were dispersed, random, or clustered pattern.
occurrences of the data. This test generates a report of the data’s significance with a z-score, p-value, and Moran’s Index value. The formal equation for the Global Moran’s I, is:

\[
I = \frac{n}{S_0} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} z_i z_j}{\sum_{i=1}^{n} z_i^2}
\]

Where \( z \) is a cell’s total number of tweets deviated from the mean of all the cell’s summed tweets \( (x_{i,j} - \bar{X}) \), \( w_{i,j} \) is a spatial weights matrix between cell \( i \) and it’s neighboring cell \( j \), \( n \) is the total number of cells, and \( S_0 \) is the aggregate of all of the spatial weights. Hence, highly clustered data should produce a small Moran’s Index value and high z-score.

The ability for users with smartphones to share their location specific coordinates, makes this test appropriate for understanding dynamics of large-scale episodic populations. These populations follow some form of a timeline as kickoff nears and games end, but are unique in where they actually spend their time both close and distant to the event venue. By simultaneously evaluating locations and tweet occurrences, a spatial correlation assessment can be made for finding the most relevant search criteria’s of game day fans. And, because this test evaluates the entire data and its patterns, improved data collection strategies can be designed appropriately when the data are proven to be not random. For example, knowing when and where to gather social media data will produce more effective results. Further, predictive measures of special event population trends can be possibly extended to future games, or even similar special events. Table 2 reports the z-score, p-value, and Moran’s Index value for each of the Top 7 terms for those tweets that fall within a 510 mile radius of Neyland Stadium (for maps of the Top 7 Terms’ distributions at all scales see APPENDIX F, G, H). This distance was chosen to include all of the season’s opponents, with the exception of the University of Oregon, due to its extreme distance from the University of Tennessee’s normal opponents.

Overall, the results were varied. Those keywords initiated by the university (e.g. “gbo” and “vfl”) were low for their z-scores, meaning statistically they were created randomly within the defined buffer. This suggests a nationwide University of Tennessee fan base were tweeting with these terms and not just those fans closest to the university. The most unanticipated result is from the term “Neyland.” While this term may be ambiguous because it is a noun and nongeographic like the others, the original assumption was that these tweets should only be
Table 2. Global Moran’s $I$ summary of each search term’s associated tweets, as well as the combined total of all tweets.

<table>
<thead>
<tr>
<th>Search Term</th>
<th>z-score</th>
<th>p-value</th>
<th>Moran's Index</th>
<th>$H_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Football Time In Tennessee&quot;</td>
<td>2.934551</td>
<td>0.003340</td>
<td>0.010409</td>
<td>Reject, clustered</td>
</tr>
<tr>
<td>&quot;Tennessee&quot;</td>
<td>4.496090</td>
<td>0.000007</td>
<td>0.002357</td>
<td>Reject, clustered</td>
</tr>
<tr>
<td>&quot;Vols&quot;</td>
<td>0.049643</td>
<td>0.960407</td>
<td>-0.001489</td>
<td>Accept, random</td>
</tr>
<tr>
<td>&quot;GBO&quot;</td>
<td>-1.135015</td>
<td>0.256369</td>
<td>-0.001566</td>
<td>Accept, random</td>
</tr>
<tr>
<td>&quot;VFL&quot;</td>
<td>-1.023034</td>
<td>0.306292</td>
<td>-0.002121</td>
<td>Accept, random</td>
</tr>
<tr>
<td>&quot;Big Orange&quot;</td>
<td>0.141479</td>
<td>0.888281</td>
<td>-0.002966</td>
<td>Accept, random</td>
</tr>
<tr>
<td>&quot;Neyland&quot;</td>
<td>-0.749570</td>
<td>0.453514</td>
<td>-0.005149</td>
<td>Accept, random</td>
</tr>
<tr>
<td><strong>Top 7 COMBINED</strong></td>
<td>9.8119</td>
<td>0.000000</td>
<td>0.0022</td>
<td>Reject, clustered</td>
</tr>
</tbody>
</table>

found surrounding Neyland Stadium; thus showing clustering. This was not observed in the data and may suggest that these users were not actually at the stadium, but still referenced the geographic location of the event in their tweets. On the contrary, the Facebook check-in data for Neyland Stadium can fill this void of understanding local episodic population dynamics that are not obvious on Twitter.

For the Top 7 combined, however, a Moran’s Index of 0.002 and a z-score of 9.8 were calculated, suggesting the data as a whole were highly clustered. The p-value was above a 99% significance level also rejecting the null hypothesis that there is no randomness to these tweets when looked at altogether. This suggests that the social media data should be looked at as a combined input to explain these special event population dynamics. When analyzing these terms individually, outside factors begin to diminish the usefulness of individual terms and instead point out the ambiguities that comprise them. In other words, these Top 7 terms complement one
another and build a stronger suited special event population signature and should be used together for statistically significant insights into the population associated with university events.

When analyzed at multiple scales, the data again proved that as a whole, tweets were statistically useful for this research. In (Table 3), all scales were significantly clustered with the exception of five miles from Neyland Stadium. At 1.5 miles, it was assumed that the results would be absolutely clustered simply because of the particular influence the football games had on campus and the directly surrounding areas. At mile five however, the z-score of 0.0815 indicates that the tweets are no different than random chance. This is surprising considering again, it is the next smallest scale to Neyland Stadium, but also because all of the other buffers suggest clustering. One explanation of this is that of the tweets that fall within this five mile buffer, less than 10% are actually outside of the 1.5 mile buffer. This lack may be caused by the many neighborhoods located outside the 1.5 mile buffer where there are more likely individual fans, instead of public gathering areas like restaurants where multiple fans may go to watch the game.

The case where Global Moran’s I was most varied was when home and away games were analyzed. In (Table 4), home and away games’ tweets were analyzed at each distance buffer to evaluate when Twitter was most significant. For those tweets that occurred during home games,

<table>
<thead>
<tr>
<th></th>
<th>z-score</th>
<th>p-value</th>
<th>Moran's Index</th>
<th>H₀</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.5 miles</td>
<td>4.4165</td>
<td>0.0000</td>
<td>0.0163</td>
<td>reject</td>
</tr>
<tr>
<td>5 miles</td>
<td>0.0815</td>
<td>0.9350</td>
<td>-0.0013</td>
<td>ACCEPT</td>
</tr>
<tr>
<td>10 miles</td>
<td>2.6965</td>
<td>0.0070</td>
<td>0.0047</td>
<td>reject</td>
</tr>
<tr>
<td>20 miles</td>
<td>4.1977</td>
<td>0.0000</td>
<td>0.0041</td>
<td>reject</td>
</tr>
<tr>
<td>50 miles</td>
<td>2.2601</td>
<td>0.0238</td>
<td>0.0012</td>
<td>reject</td>
</tr>
<tr>
<td>510 miles</td>
<td>9.8119</td>
<td>0.0000</td>
<td>0.0021</td>
<td>reject</td>
</tr>
</tbody>
</table>
Table 4. Global Moran’s I report of spatial autocorrelation for the combined Top 7 Terms used at Home or Away games, for multiple scales.

<table>
<thead>
<tr>
<th></th>
<th>HOME z-score</th>
<th>AWAY z-score</th>
<th>HOME p-value</th>
<th>AWAY p-value</th>
<th>HOME Moran’s I</th>
<th>AWAY Moran’s I</th>
<th>HOME $H_0$</th>
<th>AWAY $H_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.5 miles</td>
<td>4.2753</td>
<td>0.8213</td>
<td>0.0000</td>
<td>0.4115</td>
<td>0.0165</td>
<td>0.0084</td>
<td>reject</td>
<td>accept</td>
</tr>
<tr>
<td>5 miles</td>
<td>-0.6028</td>
<td>0.6212</td>
<td>0.5466</td>
<td>0.5345</td>
<td>-0.0033</td>
<td>-0.0006</td>
<td>accept</td>
<td>accept</td>
</tr>
<tr>
<td>10 miles</td>
<td>0.4451</td>
<td>2.3983</td>
<td>0.6566</td>
<td>0.0165</td>
<td>-0.0004</td>
<td>0.0091</td>
<td>accept</td>
<td>reject</td>
</tr>
<tr>
<td>20 miles</td>
<td>0.5584</td>
<td>5.6986</td>
<td>0.5766</td>
<td>0.0000</td>
<td>-0.0004</td>
<td>0.0223</td>
<td>accept</td>
<td>reject</td>
</tr>
<tr>
<td>50 miles</td>
<td>0.7120</td>
<td>3.4655</td>
<td>0.4765</td>
<td>0.0005</td>
<td>0.0000</td>
<td>0.0042</td>
<td>accept</td>
<td>reject</td>
</tr>
<tr>
<td>510 miles</td>
<td>4.9265</td>
<td>3.2323</td>
<td>0.0000</td>
<td>0.0012</td>
<td>0.0014</td>
<td>0.0021</td>
<td>reject</td>
<td>reject</td>
</tr>
</tbody>
</table>

Clustering was only found at the smallest (1.5 mi) and largest (510 mi) buffers. This is different from the previous table in that when away tweets are removed, the remaining buffers are no different than random. Again, the 1.5 mile buffer is not surprising. For the remaining buffers though, an explanation of this clustering at 510 and non-clustering for the others may be the results of no opponent stadium’s being located within 50 miles of Neyland Stadium (see Appendix H.2). At 510 miles, a large portion of the country is located here (see Appendix H.1), and naturally, major cities and opponent college towns are probably what make home game tweets again declare significant clustering.

The away games’ tweets are also varied and only showed clustering for areas 5 miles outside of Neyland Stadium. While 45% of these tweets occurred on away games, less than 10% were located within 5 miles. This would suggest that fans do not come to the campus area to watch the away games and instead possibly go to local eateries or friend’s homes for the comradery. Again, at 510 miles though, the study area may be too large and simply reflect the geographic component of city clustering.
Local Clustering

Because this research aims to create a model of high-resolution population distributions for Special Events, the next step examined the local clustering of tweets using ArcMap’s Cluster and Outlier Analysis tool to find the areas of high or low importance. This tool uses the Anselin Local Moran’s I test (Anselin 1995) which analyzes each feature to its neighboring features based on attributes associated with those features. In this case, the tweets captured within the cell size of 3 arc seconds are compared and a statistical report, including a z-score, p-value, Local I value, and COType (cluster/outlier), are produced for each individual cell. The equation for Anselin Local Moran’s I test is as follows:

\[ I_s = \frac{(x_i - \bar{X}^2)}{S_i^2} \sum_{j=1, j\neq i}^{n} w_{i,j}(x_i - \bar{X}^2) \]

Where, \( x_i \) is a cell’s total number of tweets, \( \bar{X} \) is the mean of all of the cells’ summed tweets, \( w_{i,j} \) is a spatial weights matrix between cell \( i \) and it’s neighboring cell \( j \), \( n \) is the total number of cells, and \( S \) is:

\[ S_i^2 = \frac{\sum_{j=1, j\neq i}^{n}(x_i - \bar{X}^2)}{n - 1} - \bar{X}^2 \]

Using at least one neighboring cell, each feature was given a COType of HH, LL, HL, LH if the feature had at least a 95% confidence. These four categories label individual features as either: a statistically significant high value surrounded by similarly high valued neighbors (HH), vice versa with low significance and neighbors (LL), a statistically high value surrounded by low values (HL), or a statistically low value surrounded by high values (LH). Out of the nearly 14,000 features (or cells in this case), 349 produced a COType classification, and only HH and HL were provided given the 95% confidence requirement. The remaining were classified “not significant.” Figure 9 displays these results surrounding Neyland Stadium.

The cells in orange are the cells with the highest relevance for game day tweets. They are the result of a statistically significant cell surrounded by statistically significant cells (HH). These areas that surround the stadium, student housing, and dining locations (such as bars and restaurants) near The Strip on Cumberland Avenue and Market Square in downtown Knoxville, are the most significant areas for game day clustering and where much of the attention should be
focused when modeling episodic populations. Therefore, these areas will be compared to the Special Events output discussed in the next section (see Special Events Model).

**Special Events Model**

As mentioned previously, the main purpose of this research was to create a model, like that of ORNL’s LandScan USA, to represent population dynamics of a special event in high-resolution distributions. Based on the analysis discussed in this paper, a Special Event distribution model was created for the University of Tennessee’s expected distribution of fans for football game days. Two outputs are presented displaying two separate times during a UT football game day when fans are not in the stadium before the game begins, and when the game is in session and all fans are assumed to have stopped moving for the duration of the game. Both distributions are in the 1.5 mile study area around Neyland Stadium. The LandScan USA nighttime dataset was used.
to build upon, as well as compare against because it best represented the study area when no
events were occurring. It was used instead of the daytime baseline for the reason that these
particular events occur on the weekend, and LandScan USA Day is a diurnal distribution model
of an average weekday, which includes worker and school populations (as previously mentioned
in Chapter 2).

With social media being an additional source of input data, this section, explains how in a
bottom-up approach, the Special Event Dynamics model was modified. From the tweets and
check-ins as weighted distribution rasters, to the assumed game day population total, this model
is broken into two parts representing a 2.5 hour pre-game scenario and a 30 minutes post-kickoff
game time scenario. The first scenario represents an empty stadium and local population
distributed in and around the campus. The second scenario describes peak population in the
stadium.

"Before The Game” Model

To establish a population count representative of the number of fans before a game, two
populations were administered that account for all fans attending and not attending the game.
These two variables include the average attendance number recorded for the season ($P_A$), which
was 95,986, and an assumed number of fans that participated in the study area, but did not attend
the game ($\alpha$) (e.g. worker population, fans without tickets). Alpha is used for the latter
population to allow for multiple scenarios of what this non-game attending population may be.
Because this population has not yet been studied, this research applied a count of 30,000 fans.
However, this count is not fully assumed for the “before game” model, only for the “during
game” model. Accordingly, a fraction of this total was used in this model to account for those
populations that have not yet arrived to campus, and is represented by $\lambda$. For this research, $\lambda$ was
arbitrarily chosen as two-thirds. Therefore, the formulation for the total population before game
time ($P_{before}$) is:

$$P_{before} = \lambda(P_A + \alpha)$$
This total of 83,991 was then distributed throughout the 1.5 mile study area by multiplying, a weight \((w)\) per cell \((i)\) based on a ratio to the sum of all the weights, thus creating a special event dataset \((P_{igameday})\):

\[
P_{igameday} = (P_{before}) \left(\frac{w_i}{\sum w_i}\right)
\]

Where, \(w_i\) is the outcome of two input density rasters added together based on the tweets \((T)\) and check-ins \((F)\) collected in the hours \((h)\) outside two hours before to three hours following kickoff. Both input rasters were computed using ArcMap’s Kernel Density tool for tweets and check-ins, respectively (Figures 10a & 10b). This technique was implemented to provide a weighted surface raster that decreased in weight as the distance increased from an observed tweet or check-in. Considering the social media data were samples of the population attending the games, a surface raster was needed to fill in the areas where populations may have been, but were not observed.

\[
w_i = \sum_{h=-12}^{12} (T_{ih})(F_{ih}) - \sum_{h=-2}^{3} (T_{ih})(F_{ih})
\]

Thus, the special event dynamics model for 2.5 hours prior to kickoff was equated at cell level \((Special Event Model_{ibefore})\), using LandScan USA’s nighttime baseline dataset \((P_{ibaseline})\) plus the updated special event dataset \((P_{igameday})\):

\[
Special Event Model_{ibefore} = P_{ibaseline} + P_{igameday}
\]
Figures 10(a&b). Using ArcMap’s Kernel Density Tool, these figures represent the weighted input rasters of tweets (10a) and check-ins (10b) for data collected before 2 hours prior to kick off and after 3 hours following kickoff.

10a. Twitter’s weighted input raster depicting 2.5 hours before kickoff using ArcMap’s Kernel Density tool.

10b. Facebook’s weighted input raster depicting 2.5 hours before kickoff using ArcMap’s Kernel Density tool.
“During The Game” Model

While the second model advances in the exact same way as the previous model, there were two variables that changed in order to properly represent distributions 30 minutes after kickoff. The first change affected the assumed game day population total by removing λ to create a maximum population count of fans attending the game ($P_A$) and the assumed additional 30,000 fans not attending the game ($\alpha$). Therefore, the new equation for calculating the population during game time ($P_{during}$) is expressed as:

$$P_{during} = P_A + \alpha$$

This new total of 125,986 was distributed, just as before, throughout the 1.5 mile study area, by multiplying it by a weight ($w$) per cell ($i$) based on a ratio to the sum of all the weights:

$$P_{i gameday} = (P_{during}) \left( \frac{w_i}{\sum w_i} \right)$$

Where again, $w_i$, is the outcome of two input density rasters added together based on the tweets ($T$) and check-ins ($F$), but only for the data collected between two hours ($h$) before to three hours following kickoff. This second change of the models’ variables, refocused those social media data that occurred right before the game and during the game, to adjust distributions in and around the stadium. And again, both input rasters were computed using ArcMap’s Kernel Density tool for tweets and check-ins, respectively (Figures 11a & 11b).

$$w_i = \sum_{h=-2}^{3} (T_{ih} + F_{ih})$$

Thus, the special event dynamics model for 30 minutes following kickoff was equated at cell level ($Special Event Model_{iduring}$), using LandScan USA’s nighttime baseline dataset ($P_{ibaseline}$) plus the newly created special event dataset ($P_{igameday}$):

$$Special Event Model_{iduring} = P_{ibaseline} + P_{igameday}$$
11a. Twitter’s weighted input raster during game time hours using ArcMap’s Kernel Density tool.

11b. Facebook’s weighted input raster during game time hours using ArcMap’s Kernel Density tool.

Figures 11(a&b). Using ArcMap’s Kernel Density Tool, these figures represent the weighted input rasters of tweets (11a) and check-ins (11b) based on those data collected from 2 hours before to 3 hours following kickoff.
With these weighted input rasters, plus the LandScan USA nighttime dataset (Figure 12), the Special Event Model for UT’s football game day distribution, at two moments during the day, is given in Figures 13a & 13b.

Figure 12. LandScan USA nighttime baseline dataset at 1.5 miles surrounding Neyland Stadium at the University of Tennessee, Knoxville.
Figure 1. The Special Event population distribution models for the University of Tennessee’s football game days for 2.5 hours before kickoff (13a) and 30 minutes after kickoff (13b).
Figure 13a represents 2.5 hours before kickoff with almost 84,000 game day fans distributed in the defined study area. Areas with the largest populations are designated along Cumberland Ave where much of the eateries are located, throughout campus where there are many student resident halls and tailgating and/or parking lots, as well as just northeast of Neyland Stadium where Market Square is located with additional eateries. These areas have produced the majority of fans because of the many tweets and check-ins found within these cells. This model therefore depicts fans coming to the campus area for the purpose of getting food before the game, or hanging out at local hot spots to participate in pregame activities, like tailgating. Additional areas with increased population are along the river where the VolNavy is located, just south of the Tennessee River where off campus student housing is located, and throughout the rest of the study area for nearby neighborhoods and businesses.

Figure 13b represents 30 minutes after kickoff with just over 125,000 additional fans being added to the study area, and nearly 100,000 of those fans being located directly within Neyland Stadium. Because of the large capacity of Neyland Stadium, much of the populations throughout campus decreases to represent how fans left their previous location to attend the game. There remain some populations however, to suggest that not everyone who attends UT’s game day activities actually attend the game. In fact, only 12.6% of the seats in Neyland Stadium are designated for students only\(^8\), and with nearly 28,000 students enrolled\(^9\) at the university, many of these students cannot attend the game unless they buy tickets assigned to the general public. Therefore, to accommodate those students who do not attend, as well as fans who are not students and not attending the game, areas with restaurants and bar remain highly populated.

**Discussion**

Social media has many interesting informational values providing real-time updates of events around the world and has emerged as an important source of data with potential for applications in science and operations. However, there are many challenges and uncertainties

---


associated with social media data including unknown nature of the data provider, unreliability of
the information, and non-uniform contributions in space, time, and demographic sectors.
Additionally, deriving similar authoritative and statistical values of well-designed survey based
data collection approaches are still not feasible. Nevertheless, social media data still remains as
an openly accessible stream of data that holds information about dynamics of a population that is
not easily available otherwise and its potential and usability in geographic and demographic
modeling should be explored systematically.

Correlations between the social media data and reported attendance at the games were
analyzed to identify possible spatial and temporal characteristics of the data that could be
valuable for improving a special event population dynamics model. The correlations between
what was truth (i.e. attendances) with what was provided (i.e. tweets and check-ins) were
extremely low, but nevertheless slightly promising for not being collected without error and from
being produced by a small sample of the general public. Even with the unexpected collection
gaps or mishaps, the data still showed an encouraging positive correlation for tweets to
attendance and check-ins to attendance. If this research is repeated with consistent results, there
could be predictive powers that should assist in emergency response and management. When
lives are in question, low but positive results should not be ignored by the traditional
expectations of statistically significant results. The challenge instead, should be to mimic this
study for other episodic populations when local knowledge is not known.

The local knowledge for this paper demonstrated that very little actually needs to be
known for a search criteria to be made. From 35 university specific keywords used to scrape
Twitter’s API feed, only seven were sufficient when exploring social media signatures. Even
still, half of those seven were only necessary when analysis was local. The global clustering of
terms may have shown a presence of followers and fans throughout the country, but not a
significant clustering worth noting. This is nevertheless, essential in understanding from where
populations may be coming or where those displaced populations may need to go. However, it is
not paramount to first responders in near real-time. Catastrophic actions at Special Events will be
putting an immediate local gathering of individuals in danger, and this is where local clustering
analysis is what is going to matter most. The results from the Anselin Local Moran’s I test
identified the hotspots where a majority of the fan base would be found based on the captured
tweets. Social media data, therefore, provided enough important spatio-temporal characteristics for episodic populations to be represented. This type of suggestive insights is what makes this research necessary for the discipline of population modeling and what could potentially improve population distribution modeling techniques such as LandScan USA.

It is important to note that while a portion of the tweets were geolocated, the accuracy of cell phone positioning has its own limitations and granularity, and so it is helpful to use Facebook Places as an extension to model movements. But again, local knowledge is not entirely necessary. Out of 95 Facebook places in a 1.5 mile buffer from Neyland Stadium, under half were used each week. With increasingly stricter privacy limitations to Facebook on user posts, check-ins are currently the only avenue to interpreting fan based dynamics. This is still helpful when modeling the flow of these game day fans since check-ins function as a known source of check points throughout the day. Additionally, with check-ins to the stadium having the same signature for games with different start-times and opponents, it can be assumed populations move in and around the stadium at the same rate each game. This is a supportive method to prepare for future games at least 2 hours before kickoff, and should be the goal for future studies to correlate check-ins with local geocoded tweets. Or, if access to digital ticket readers is available, correlations to those records would be the best validation to check-ins.

A few unfortunate misrepresentations of social media’s geolocated data are the issues with cross posting between social media platforms, and the time lapse in actual arrivals to a venue or actual presence at a venue. Cross posting is an option offered to another social media platform, known as Instagram. This photo sharing media allows users to upload any photo and caption to the service’s own website, as well as share that same post to other social media supporting websites like, Facebook, Twitter, Tumblr, Foursquare, and Flickr. By selecting one of the aforementioned, like Twitter, a post is shared twice adding to each platforms streamed data. A positive side to this is if a user on Twitter who does not opt in to share their location through Twitter’s geolocated settings, originates a post from their Instagram account with a tagged location to their picture; through this cross posting they are in fact generating a geolocated tweet through Twitter’s API. This was the case with 638 geolocated tweets that were actually Instagram post of pictures tagged with “Neyland Stadium” that appeared on the west side of the stadium (see APPENDIX I). However, this process is also damaging to the metadata of social
media. For example, users who tag their photos through Instagram’s locations automatically generate a tweet at Instagram’s coordinates for that venue. A user does not actually have to be located where the tweet is geographically assigned. Plus, users do not necessarily always share posts at the exact moment when a picture is taken. Users can decide later that a photo is worth sharing, tag a location to it, and have it post hours after they have left that venue. These types of social media shares will skew how the data is interpret, but ultimately, still provide a locational understanding of a user even when that user may not have manually opted in to sharing through Twitter’s privacy settings.

Aside from these inconsistencies in the data providers, as well as misrepresentations of metadata, social media is a true source of ground truthing for episodic populations not normally accounted for in current population models. No longer will researchers require additional ex-post-facto reports to confirm special event population dynamics, when many individual observations are already at hand. Instead of verifying where fans gather and move to by collecting news reports, interpreting blogs, or waiting for traditional expert records to be published, social media gives insight into these dynamics immediately. This research proved that even fewer search criteria’s or parameters are needed to collect sufficient amounts of data to prove where significant clusters of people are located and when. With a little refinement, and better collecting processes, this research could produce even finer representations of special event populations. For now however, this research successfully explored the feasibility of using social media as an additional source to modeling a special event distribution dataset.
CHAPTER V
CONCLUSION

This paper examined the feasibility of incorporating social media data during the 2013-2014 football season of the University of Tennessee to enhance the understanding of population distribution and dynamics during episodic events. Using tweets and check-ins, analysis was performed on social media’s two most popular platforms, Twitter and Facebook, for creating predictive signatures and spatial patterns. Keywords associated with the university were identified to mine Twitter’s API, and over 2 million tweets were returned. Eventually, a list of the Top 7 terms were defined to create the most representative activity signatures with still over 1.7 million tweets. Additionally, 95 Facebook Places were also identified within 1.5 miles of Neyland Stadium in order to support location of fan based hot spots, but only 43 were ever used. From these data, many measurable and/or predictable structures to game day social media dynamics were concluded, and as such a special event distribution model was produced for two significant moments during the an average football game day.

While statistical results were not highly significant in predicting social media data based solely on attendance, the results were still slightly positive and are potential indicators that there may be some correlation and should be examined further. Had all the tweets been collected properly (i.e. consistently, and with the proper weekly opponents’ terms), and if all of the games had reported some check-in data (i.e. for Games 10-12), there might have been a higher significant correlation in social media prediction. Too few observations hindered this portion of the research which required additional analysis to test the significance of the observed social media data.

The spatial analyses performed examined the relationships of tweets separately and collectively, at multiple scales, and for different games (home or away) to indicate geographic relevance associated with game day dynamics. For the Global Moran’s $I$ test, tweets proved most effective when analyzed together or when the study area was small or extremely large. Areas that were of the highest significance to local game day fans were identified at the LandScan USA cell size of 3 arc seconds using tweets from the LISA test. With confidence being supported by these two tests, additional weighted input rasters were created using the raw social media data, and a linear regression distance model to create a high-resolution distribution dataset of UT football
games day fans. At such a fine resolution, this model is the best representation for the areas surrounding the University of Tennessee’s campus, and should be the model used for emergency management and response analysis. Furthermore, social media data proved to be very successful as an additional source to population modeling, and should be an accepted form of input data to model episodic populations.

There are many ways this research could have been improved and some are suggested here for future work. For starters, data should be collected consistently. Results would change dramatically if each and every game were collected, as well as collected using the proper week’s terms. With that said, not all terms associated need to be implemented. This can allow for more returned tweets since the API only allows a small sample; however, initial requests of many search terms may actually be necessary to determine those most efficient keywords. Additionally, multiple platforms should be considered. Facebook is still a re-enforcing indicator of venues that support tailgating/hang out spots before and after game, though may not be used as much as Twitter. Instagram is an increasingly popular media that is becoming widely used to share information in the social media world. In fact, some posts are actually initiated through Instagram and then shared to other platforms. As discussed previously, this cross platform sharing can alter the geolocated observations that define the weighted rasters thus changing the output distribution. Additionally, examining the capabilities of Instagram, or any other platform for that matter, could potentially be a work around or a data gap filler to the restrictions found on Twitter or Facebook.

Until researchers can have full access to social media, population modelers must accept that these observations from public users are the best and quickest form of ground truth. Using volunteered information as a source of live updates is a growing interest for those interested in Big Data, and it is clear as to why. Data collection is expedited and can be done with little or no cost. Therefore, it should begin to become an accepted research ideology among the sciences. And because of its untraditional form of information retrieval, less traditional expectations must be expected of what is significant and find a threshold that meets both the sciences’ requirements and realizations that time is of the essence in emergency management and response. Hopefully, this paper proves the growing feasibility of using non-traditional sources, such as social media, as a form of validation and verification for near real-time population modeling techniques.


Kautz, H. 2013. Data Mining Social Media for Public Health Applications. 23rd International Joint Conference on Artificial Intelligence (IJCAI 2013), Beijing, China, 2013.


APPENDIX
APPENDIX A

Examples of Raw, streamed data (pages 60-61)
### Appendix A.1

A snapshot of Twitter’s raw, streamed data for the top 35 university specific terms, in total counts every 15 minutes.

|----------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|------------- |
### Appendix A.2

A snapshot of Facebook’s raw, streamed data for 95 Facebook Places, every 30 minutes.

<table>
<thead>
<tr>
<th>Place Name</th>
<th>Aug 30, 2013</th>
<th>8/31/2013 16:30</th>
<th>8/31/2013 17:00</th>
<th>8/31/2013 17:30</th>
<th>8/31/2013 18:00</th>
<th>8/31/2013 18:30</th>
<th>8/31/2013 19:00</th>
<th>8/31/2013 19:30</th>
<th>Dec 1, 2013</th>
</tr>
</thead>
<tbody>
<tr>
<td>17th Street Market &amp; Deli</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>Blount Mansion</td>
<td>67</td>
<td>67</td>
<td>67</td>
<td>67</td>
<td>67</td>
<td>67</td>
<td>67</td>
<td>67</td>
<td>67</td>
</tr>
<tr>
<td>Buffalo Wild Wings</td>
<td>35</td>
<td>35</td>
<td>35</td>
<td>35</td>
<td>35</td>
<td>35</td>
<td>35</td>
<td>35</td>
<td>35</td>
</tr>
<tr>
<td>Cafe 4</td>
<td>3439</td>
<td>3440</td>
<td>3440</td>
<td>3440</td>
<td>3440</td>
<td>3440</td>
<td>3440</td>
<td>3440</td>
<td>3440</td>
</tr>
<tr>
<td>Cairo Cafe Mirage</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Calhoun’s Restaurants</td>
<td>4625</td>
<td>4626</td>
<td>4627</td>
<td>4627</td>
<td>4627</td>
<td>4627</td>
<td>4627</td>
<td>4627</td>
<td>4627</td>
</tr>
<tr>
<td>Chesapeake’s</td>
<td>1552</td>
<td>1552</td>
<td>1553</td>
<td>1553</td>
<td>1553</td>
<td>1553</td>
<td>1553</td>
<td>1553</td>
<td>1553</td>
</tr>
<tr>
<td>Chipotle Mexican Grill</td>
<td>141</td>
<td>141</td>
<td>141</td>
<td>141</td>
<td>141</td>
<td>141</td>
<td>141</td>
<td>141</td>
<td>141</td>
</tr>
<tr>
<td>Circle Park</td>
<td>224</td>
<td>224</td>
<td>224</td>
<td>224</td>
<td>224</td>
<td>224</td>
<td>224</td>
<td>224</td>
<td>224</td>
</tr>
<tr>
<td>Coffee and Chocolate</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Cool Beans Bar and Grill</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Copper Cellar Restaurants</td>
<td>1256</td>
<td>1256</td>
<td>1256</td>
<td>1257</td>
<td>1257</td>
<td>1257</td>
<td>1257</td>
<td>1257</td>
<td>1257</td>
</tr>
<tr>
<td>Cumberland Grill</td>
<td>36</td>
<td>36</td>
<td>36</td>
<td>36</td>
<td>36</td>
<td>36</td>
<td>36</td>
<td>36</td>
<td>36</td>
</tr>
<tr>
<td>Dazzo’s Pizzeria</td>
<td>38</td>
<td>38</td>
<td>38</td>
<td>38</td>
<td>38</td>
<td>38</td>
<td>38</td>
<td>38</td>
<td>38</td>
</tr>
<tr>
<td>Domino’s Pizza UT Knoxville</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Downtown Grill &amp; Brewery</td>
<td>11657</td>
<td>11659</td>
<td>11659</td>
<td>11660</td>
<td>11661</td>
<td>11661</td>
<td>11661</td>
<td>11661</td>
<td>11661</td>
</tr>
<tr>
<td>Ephesus Mediterranean Grill</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>Firehouse Subs Cumberland Ave</td>
<td>395</td>
<td>395</td>
<td>395</td>
<td>395</td>
<td>395</td>
<td>395</td>
<td>395</td>
<td>395</td>
<td>395</td>
</tr>
<tr>
<td>Fort Dickerson Park</td>
<td>45</td>
<td>45</td>
<td>45</td>
<td>45</td>
<td>45</td>
<td>45</td>
<td>45</td>
<td>45</td>
<td>45</td>
</tr>
<tr>
<td>Fort Kid</td>
<td>39</td>
<td>39</td>
<td>39</td>
<td>39</td>
<td>39</td>
<td>39</td>
<td>39</td>
<td>39</td>
<td>39</td>
</tr>
<tr>
<td>Fort Sanders Yacht Club</td>
<td>216</td>
<td>216</td>
<td>216</td>
<td>216</td>
<td>216</td>
<td>216</td>
<td>216</td>
<td>216</td>
<td>216</td>
</tr>
<tr>
<td>Foundry On The Fair Site</td>
<td>156</td>
<td>156</td>
<td>156</td>
<td>156</td>
<td>156</td>
<td>156</td>
<td>156</td>
<td>156</td>
<td>156</td>
</tr>
<tr>
<td>Four Points by Sheraton Knoxville Cumberland House</td>
<td>537</td>
<td>537</td>
<td>537</td>
<td>537</td>
<td>537</td>
<td>537</td>
<td>537</td>
<td>537</td>
<td>537</td>
</tr>
<tr>
<td>G10 Parking Lot</td>
<td>80</td>
<td>80</td>
<td>81</td>
<td>81</td>
<td>81</td>
<td>81</td>
<td>81</td>
<td>81</td>
<td>81</td>
</tr>
<tr>
<td>Gus’s Good Times Deli</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Gyrene Burger</td>
<td>35</td>
<td>35</td>
<td>35</td>
<td>35</td>
<td>35</td>
<td>35</td>
<td>35</td>
<td>35</td>
<td>35</td>
</tr>
<tr>
<td>Hampton Inn &amp; Suites Downtown Knoxville</td>
<td>552</td>
<td>552</td>
<td>552</td>
<td>552</td>
<td>552</td>
<td>552</td>
<td>552</td>
<td>552</td>
<td>552</td>
</tr>
<tr>
<td>Hanna’s cafe</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Hibachi Factory Cumberland Ave</td>
<td>104</td>
<td>104</td>
<td>104</td>
<td>104</td>
<td>104</td>
<td>104</td>
<td>104</td>
<td>104</td>
<td>104</td>
</tr>
<tr>
<td>Hilton Knoxville</td>
<td>2960</td>
<td>2960</td>
<td>2960</td>
<td>2960</td>
<td>2960</td>
<td>2960</td>
<td>2960</td>
<td>2960</td>
<td>2960</td>
</tr>
<tr>
<td>Holiday Inn World’s Fair Park</td>
<td>620</td>
<td>620</td>
<td>620</td>
<td>620</td>
<td>620</td>
<td>620</td>
<td>620</td>
<td>620</td>
<td>620</td>
</tr>
<tr>
<td>House of Dragon</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>International House</td>
<td>19</td>
<td>19</td>
<td>19</td>
<td>19</td>
<td>19</td>
<td>19</td>
<td>19</td>
<td>19</td>
<td>19</td>
</tr>
<tr>
<td>Jal Dee Thai &amp; Japanese Cuisine</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Jimmy John’s</td>
<td>31</td>
<td>31</td>
<td>31</td>
<td>31</td>
<td>31</td>
<td>31</td>
<td>31</td>
<td>31</td>
<td>31</td>
</tr>
<tr>
<td>Knox County Courthouse</td>
<td>760</td>
<td>760</td>
<td>760</td>
<td>760</td>
<td>760</td>
<td>760</td>
<td>760</td>
<td>760</td>
<td>760</td>
</tr>
<tr>
<td>Krystal</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>Laurel Mountain Eatery</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>Longbranch Saloon</td>
<td>458</td>
<td>458</td>
<td>458</td>
<td>458</td>
<td>458</td>
<td>458</td>
<td>458</td>
<td>458</td>
<td>458</td>
</tr>
<tr>
<td>Market Square, Knoxville</td>
<td>432</td>
<td>432</td>
<td>432</td>
<td>432</td>
<td>432</td>
<td>432</td>
<td>432</td>
<td>432</td>
<td>432</td>
</tr>
</tbody>
</table>
APPENDIX B

Individual Game Day Twitter Activity:
24 Hours Surrounding Kickoff (pages 63-70)
Appendix B.1

Game 1 – vs Austin Peay at 6:00pm

- Football time in Tennessee
- Tennessee
- Vols
- GBO
- VFL
- Big Orange
- Neyland
- APSU
- Austin Peay
- Warriors

Kickoff
Appendix B.3

Game 3 – vs Oregon at 3:30pm

Kickoff
Appendix B.5

Game 7 – vs South Carolina at 12:00pm

Kickoff

football time in tennessee
tennessee
vols
gbo
vfl
big orange
neyland
gamecocks
south carolina
gococks
spurrier
visor
usc
APPENDIX C

Facebook Check-in Activity, 24 Hours Surrounding Kickoff, For Each Home Game (pages 72-76)
Appendix C.1

Total game day check-ins for Game 1, vs. APSU on August 31\textsuperscript{st}, 2013, for 42 out of 95 Facebook Places 1.5 miles surrounding Neyland Stadium.
Appendix C.2

Total game day check-ins for Game 2, vs. WKU on September 7\textsuperscript{th}, 2013, for 36 out of 95 Facebook Places 1.5 miles surrounding Neyland Stadium.
Appendix C.3

Total game day check-ins for Game 5, vs. USA on September 28th, 2013, for 35 out of 95 Facebook Places 1.5 miles surrounding Neyland Stadium.
Appendix C.4

Total game day check-ins for Game 6, vs. UGA on October 5th, 2013, for 37 out of 95 Facebook Places 1.5 miles surrounding Neyland Stadium.
Appendix C.5

Total game day check-ins for Game 7, vs. USC on October 19\textsuperscript{th}, 2013, for 22 out of 95 Facebook Places 1.5 miles surrounding Neyland Stadium.
APPENDIX D

Facebook’s 95 Places within 1.5 miles of Neyland Stadium that are of interest to fans attending a football game.
### APPENDIX E

Keywords used in search criteria for Twitter’s Streaming API  
(UTK baseline terms + weekly opponent’s terms)

<table>
<thead>
<tr>
<th>Vols / govols</th>
<th>GBO / gobigorange / big orange</th>
<th>smokey</th>
<th>pregame</th>
</tr>
</thead>
<tbody>
<tr>
<td>volunteers</td>
<td>tennessee</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VFL / vol for life</td>
<td>football time in tennessee</td>
<td>run through</td>
<td>tailgate</td>
</tr>
<tr>
<td>vol walk / volwalk</td>
<td>footballtimeintennessee</td>
<td>band</td>
<td></td>
</tr>
<tr>
<td>volnation</td>
<td>utk* / #utk</td>
<td>orange</td>
<td>circle park</td>
</tr>
<tr>
<td>vol navy</td>
<td>utcoachjones</td>
<td>risetothetop</td>
<td>torchbearer</td>
</tr>
<tr>
<td>vol village</td>
<td>butch</td>
<td>brickbybrick</td>
<td></td>
</tr>
<tr>
<td></td>
<td>team117</td>
<td>rocky top / rockytop</td>
<td></td>
</tr>
</tbody>
</table>

+ **Weekly Opponent’s Terms**

<table>
<thead>
<tr>
<th>Game 1 Austin Peay</th>
<th>Game 2 Western Kentucky</th>
<th>Game 3 Oregon</th>
<th>Game 4 Florida</th>
<th>Game 7 South Carolina</th>
<th>Game 8 Alabama</th>
<th>Game 9 Missouri</th>
<th>Game 10 Auburn</th>
<th>Games**: 5, 6, 11, 12 (1 mile around Neyland Stadium)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6:00</td>
<td>12:21</td>
<td>3:30</td>
<td>3:30</td>
<td>10:19/2013</td>
<td>3:30</td>
<td>7:00</td>
<td>12:00</td>
<td>South Alabama</td>
</tr>
<tr>
<td>apsu</td>
<td>wku</td>
<td>(held over)</td>
<td>gators</td>
<td>gamecocks</td>
<td>alabama</td>
<td>(held over)</td>
<td>(held over)</td>
<td>Georgia</td>
</tr>
<tr>
<td>austin peay</td>
<td>kentucky</td>
<td>from</td>
<td>florida</td>
<td>carolina</td>
<td>(roll) tide</td>
<td>from</td>
<td>from</td>
<td>231</td>
</tr>
<tr>
<td>warriors</td>
<td>petrino</td>
<td>Western</td>
<td>beatgators</td>
<td>spurrier</td>
<td>saban</td>
<td>Alabama</td>
<td>Alabama</td>
<td>Vanderbilt</td>
</tr>
<tr>
<td>hilltoppers</td>
<td>Kentucky</td>
<td>gogators</td>
<td>visor</td>
<td>crimson</td>
<td></td>
<td></td>
<td></td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>theswamp</td>
<td>gococks</td>
<td>beatbama</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Kentucky</td>
</tr>
<tr>
<td></td>
<td>muschamp</td>
<td>usc</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>15</td>
</tr>
<tr>
<td>Total:</td>
<td>Total:</td>
<td>Total:</td>
<td>Total:</td>
<td>Total:</td>
<td>Total:</td>
<td>Total:</td>
<td>Total:</td>
<td></td>
</tr>
<tr>
<td>223,782</td>
<td>219,419</td>
<td>295,904</td>
<td>248,691</td>
<td>393,855</td>
<td>370,245</td>
<td>108,855</td>
<td>216,124</td>
<td></td>
</tr>
</tbody>
</table>

**Top 7 Terms**

<table>
<thead>
<tr>
<th></th>
<th>Total:</th>
<th>Total:</th>
<th>Total:</th>
<th>Total:</th>
<th>Total:</th>
<th>Total:</th>
<th>Total:</th>
<th>Total:</th>
</tr>
</thead>
<tbody>
<tr>
<td>167,913</td>
<td>166,653</td>
<td>226,479</td>
<td>182,846</td>
<td>254,942</td>
<td>177,003</td>
<td>83,270</td>
<td>149,033</td>
<td></td>
</tr>
</tbody>
</table>

* used “utk” until it was replaced with “#utk” starting at Game 3 because of Indonesian word “for”

**not retrieved through Twitter’s API. Manual collection
APPENDIX F

All Games’ Tweets
at Multiple Scales (pages 80-85)
Appendix F.1

510 Mile Buffer

- Neyland Stadium
- Opponent Stadiums
  - 1 – Austin Peay
  - 2 – Western Kentucky
  - 3 – Oregon
  - 4 – Florida
  - 7 – South Carolina
  - 8 – Alabama
  - 9 – Mizzou
  - 10 – Auburn
Appendix F.2

50 Mile Buffer

Neyland Stadium
Opponent Stadiums
1 – Austin Peay
2 – Western Kentucky
3 – Oregon
4 – Florida
7 – South Carolina
8 – Alabama
9 – Mizzou
10 – Auburn
Appendix F.3

20 Mile Buffer

Legend:
- Neyland Stadium
- Opponent Stadiums
- 1 – Austin Peay
- 2 – Western Kentucky
- 3 – Oregon
- 4 – Florida
- 7 – South Carolina
- 8 – Alabama
- 9 – Mizzou
- 10 – Auburn
Appendix F.4

10 Mile Buffer

Neyland Stadium
- Opponent Stadiums
- 1 – Austin Peay
- 2 – Western Kentucky
- 3 – Oregon
- 4 – Florida
- 7 – South Carolina
- 8 – Alabama
- 9 – Mizzou
- 10 – Auburn
Appendix F.5

![5 Mile Buffer Map]
Appendix F.6
APPENDIX G

Home Games’ Tweets
at Multiple Scales (pages 87-92)
Appendix G.1

510 Mile Buffer

Neyland Stadium
Opponent Stadiums
1 – Austin Peay
2 – Western Kentucky
7 – South Carolina
10 – Auburn
Appendix G.2

50 Mile Buffer

Neyland Stadium
- Opponent Stadiums
  - 1 – Austin Peay
  - 2 – Western Kentucky
  - 7 – South Carolina
  - 10 – Auburn
Appendix G.3

20 Mile Buffer

- Neyland Stadium
- Opponent Stadiums
  - 1 – Austin Peay
  - 2 – Western Kentucky
  - 7 – South Carolina
  - 10 – Auburn

N

4 miles
8 miles
16 miles

Knox

Anderson

Seyler

Loudon

Blount

Union

Grainger

Jefferson

North Carolina
Appendix G.5

5 Mile Buffer

Neyland Stadium
- Opponent Stadiums
  1 – Austin Peay
  2 – Western Kentucky
  7 – South Carolina
  10 – Auburn
Appendix G.6

1.5 Mile Buffer

- Neyland Stadium
- Opponent Stadiums
  - 1 – Austin Peay
  - 2 – Western Kentucky
  - 7 – South Carolina
  - 10 – Auburn

Tennessee River
APPENDIX H

Home vs Away Games’ Tweets at Multiple Scales (pages 94-99)
Appendix H.2
Appendix H.3

20 Mile Buffer

Tweets
- Neyland Stadium
- Opponent Stadiums
- Home
- Away
Appendix H.4

10 Mile Buffer

Tweets

🌟 Neyland Stadium
🏠 Opponent Stadiums
_PURPLE: Home
 Antworten: Away

97
APPENDIX I

A snapshot of the raw geolocated tweets (red) surrounding Neyland Stadium with aggregated cell totals of those raw geolocated tweets (yellow) per 3 arc-second cells snapped to LandScan’s USA raster. The cell on the west side of the stadium (with a total of 711 tweets) includes the cross posting shares from Instagram’s tagged photos of “Neyland Stadium.”
VITA
Kelly was born in 1987 to Thomas and Deborah Sims. For 16 years, she lived in Tampa, Florida with her parents and two older siblings, Jeffrey and Valerie Pendino. On New Years Eve 2002, she moved with her parents to Knoxville, Tennessee to live the “lake life” on Norris Lake. She graduated from Gibbs High School in 2005 and started college that following Fall at the University of Tennessee. While in college she majored in Geography and minored in Tourism & Hospitality Management. In her final semester, Kelly accepted an internship with Oak Ridge National Laboratory (ORNL) in Oak Ridge, TN, thanks to her professor, Dr. Bruce Ralston, for putting her in contact with Eddie Bright in the Geographic Information Science & Technology (GIST) group. After graduating in the Spring of 2009, she was offered a full-time position with that same group as a Research Associate. For the next 3 years, Kelly worked with many highly recognized research scientists in the GeoSpatial and GeoIntelligence fields, presented at the Association of American Geographers (AAG) Annual Meeting each year, and published her first co-authored paper to *Natural Hazards* with a few of her co-workers.

In the Fall of 2012, Kelly returned to the University of Tennessee to pursue her Master’s degree. With an interest in Geographic Information Systems (GIS), as well as tourism, she focused her research around the impacts and dynamics of tourist/episodic populations. With guidance from her advisor, Dr. Budhendra Bhaduri (who is also the Group Leader for the GIST group), her thesis was constructed to model populations at risk during popular episodic events. Coincidentally, because a top SEC football team was right in her backyard, she decided to study the distribution patterns of football fans through mining social media data during the University of Tennessee’s home football games. Kelly presented some of her preliminary findings from this paper at, again, the AAG Annual Meeting, this time back in her hometown of Tampa, Florida. Following graduation she hopes to continue working as a population modeler for event risk analysis with the extremely talented people in the GIST group, spending time with her family without them telling her “just sit down and write your paper”, and enjoying her life with her longtime beau, Cory Halligan.