ASSESSING THE VALIDITY OF LOCATION-BASED SOCIAL MEDIA IN THE STUDY OF SPATIAL PROCESSES

by

MATTHEW HAFFNER

Bachelor of Science in Secondary Education
Pittsburg State University
Pittsburg, Kansas
2012

Master of Science in Geography
Oklahoma State University
Stillwater, Oklahoma
2014

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
May, 2018

ASSESSING THE VALIDITY OF LOCATION-BASED SOCIAL MEDIA IN THE STUDY OF SPATIAL PROCESSES

Dissertation Approved:

G. Allen Finchum, Ph.D.

Committee Chair

Adam J. Mathews., Ph.D.

Dissertation Adviser

Emily Fekete, Ph.D.

Sharon R. Bird, Ph.D.

Outside Committee Member

ACKNOWLEDGMENTS

I am certain that I would not have been successful in completing this dissertation and degree without the help and support of an abundance of people. It will be impossible to list them all here, but I will name a few.

Oklahoma State was a great place to pursue two graduate degrees, and I know I will remain close to many of the faculty, staff, and students I met here. I had an excellent array of fellow graduate students over the past six years, particularly Gustavo Ovando, Aswin Subanthore, Jordan Brasher, Siewe Siewe, Thomas Craig, Yun Zhao, Robert Garrett, Colton Flynn, and Nick Rose. I also had several important funding sources outside of the department during my time at Oklahoma State. I would like to thank Dr. Tyson Ochsner in the Department of Plant and Soil Science for employing me in the summer of 2015. Here I learned an incredible amount about programming, research, and academic writing. I would also like to thank Lynda Ozan and the State Historic Preservation Office in Oklahoma City for providing the funds for my research assistantship over the past two years.

I want to thank several departmental staff and faculty members. The positive attitudes of Emily Williams and Mike Larson toward students always made me feel at home in the department. Dr. Jon Comer taught one of the best courses I ever took (Geographic Analysis II) and really boosted my confidence as a quantitative researcher. Dr. Alyson Greiner also taught one of my favorite courses (History and Philosophy of Geography) and changed the way I thought about the discipline. Dr. Rebecca Sheehan's kindness toward me in my second semester as a master's student resulted in me staying in the program at a time when I was seriously considering other career paths. Dr. Brad Bays was an outstanding supervisor while I was a graduate teaching assistant, and the freedom he gave us as instructors played a pivotal role in me falling in love with teaching again. Dr. Amy Frazier and Dr. Peter Kedron were instrumental in helping me search for jobs and ultimately secure a position as an assistant professor. I am incredibly thankful for the time they put into helping graduate students (even students who are not studying under them) and for truly treating us like colleagues.

I could not have asked for a better dissertation committee. Sharon, your insights as my outside member have proved enormously beneficial, especially in improving my writing. Emily, I have greatly appreciated the standard you held me to in this dissertation and your ability in getting me to think outside of my comfortable quantitative box. I have gleaned more about geographic theory from you than anyone else.

As a part of this amazing committee, I have had an incredible advising situation. Having two co-advisors who care about me as person, take the time to mentor me, and who I can

share a good laugh with has been wonderful. Allen, I know that I learned more about geography and programming from you during conversations in your office and on rides to and from Oklahoma City than I ever learned in a classroom. I am so thankful you asked me to join you on that Twitter poster for AAG years ago, which is really what started me in this direction for my dissertation. Adam, you have always pushed me to be my best, and I would not be in the position I am in now without your guidance. Thank you for all the proofreading of my work, taking me to conferences, books on being a dad, parenting classes, beers, and everything else. I cannot thank you both enough.

My family has also been instrumental throughout this process. First, my parents were enormously supportive of my desire to pursue a PhD. Without their initial encouragement, I would not have undertaken this degree program. My siblings - Jamie, Rob, and Anna - have also been a great encouragement. To my beautiful wife, Kristin, I want to say thank you so much for your love and support over the past few years. You kept me sane, consistently reminded me what is important in life, and have been a joy to be around. I am lucky to be married to such a kind and caring person, and you are the best mom on the planet. Little Silas, the day you were born was the best day of my life, and you have motivated me in ways you cannot understand.

To close, I want to briefly discuss my motivation for pursuing the topics in this dissertation. Initially, I knew that I wanted to study big spatial data, but I was unsure how I would go about it. I ultimately decided to work toward using social media to draw out spatial patterns of inequality, address the needs of marginalized groups, and combat racism. While my ultimate inspiration lies in the life and work of Jesus Christ, several passages in the Old Testament were instrumental in shaping the way that I approached this dissertation. During the Fall of 2015, Micah 6:8, Proverbs 13:23, 14:31, 22:16, and 22:22-23 particularly influenced me, but no passage more so than Proverbs 31:8-9 which reads

> “Speak up for those who cannot speak for themselves,
>
> for the rights of all who are destitute.
>
> Speak up and judge fairly;
>
> defend the rights of the poor and needy.”

Name: MATTHEW HAFFNER

Date of Degree: MAY 2018

Title of Study: ASSESSING THE VALIDITY OF LOCATION-BASED SOCIAL MEDIA IN THE STUDY OF SPATIAL PROCESSES

Major Field: GEOGRAPHY

Abstract: The advent of big spatial data has created new opportunities for studying geographic phenomena. Open mapping projects, citizen science initiatives, and location-based social media all fall under the umbrella of volunteered geographic information and are now frequently used spatial data sources. The fact that these sources are user-contributed as opposed to gathered by experts has raised significant concerns over data quality. While data accuracy, particularly in open mapping projects (e.g., OpenStreetMap), has been given considerable attention, far less has been paid to data validity, specifically on location-based social media. In this three article dissertation, I explore the validity of location-based social media in the study of spatial processes. In the first article, I implement a survey on the Oklahoma State campus to explore college students' behaviors and perceptions of location-based social media and note differences in terms of gender, race, and academic standing. The second and third articles are empirical studies utilizing geolocated data from Twitter, a popular social media platform. The second article makes use of precise location data (e.g., latitude – longitude) and uses geographically weighted regression to explore the patterns of non-English Twitter usage in Houston, Texas. The third article uses general location data (e.g., city) to explore the patterns of #BlackLivesMatter and counter-protest content across the states of Louisiana and Texas. The results of these studies collectively provide an optimistic, though cautionary, outlook on the use of location-based social media data in geography.

TABLE OF CONTENTS

Chapter Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER I

INTRODUCTION

Recent advancements in GIScience

Plummeting storage costs, consistent improvements in computing, and the proliferation of the internet have paved the way for the fourth paradigm of science, characterized by seemingly infinitely large datasets (Elwood et al. 2013). Much of these data are spatially referenced, enabling new types of analyses within geography (Miller 2010). The study of "spatial big data" has become a major research area within geographic information science (GIScience) with emphases geared toward analyzing user-contributed sources (Egenhofer et al. 2016), termed volunteered (Goodchild 2007) or contributed (Harvey 2013) geographic information (VGI or CGI). What constitutes big data, however, is constantly changing (Graham and Shelton 2013).

Big data has traditionally been defined by the "three V's": volume (denoting large file sizes and large numbers of files), velocity (referring to the speed of data streams), and variety (meaning disparate file types and formats) (ADS 2001). Over time, this definition has become more inclusive. Data science firms now claim six (Schaafsma 2018), seven (DeVan 2016), or ten (Firican 2017) V's, with an unseemly upper limit of forty-two (Shafer 2017). Despite big data's increasingly convoluted definition, some meaningful common elements have emerged; notably, veracity and validity. While veracity pertains to data accuracy or truthfulness, validity refers to the assumptions of what the data actually measure.

The inclusion of these two new V's is significant in that it calls into question the authoritativeness of standalone data. It challenges theory-free approaches to data driven analysis, which has its proponents (e.g., Anderson 2008), and calls for more critical interpretations of new big data sources. The proliferation of user-generated content through the internet, often referred to as Web 2.0 (O'Reilly 2005), increasingly demonstrates that big data is largely socially constructed. Questions over veracity and validity force a recognition of the quality limitations in large datasets, especially in social media data such as those from Twitter. A thorough understanding of data's veracity and validity are crucial for turning big data into useful information.

Big spatial data: Boom or bust?

Despite the limitations of these new data sources, they show promise in addressing problems across a wealth of domains, especially in geography. Perhaps the largest VGI platform, OpenStreetMap (OSM), aided disaster relief efforts following the 2010 Haiti Earthquake (Zook et al. 2010), the 2011 Japan Earthquake (Imi et al. 2012), the 2012 Horsethief Canyon Fire (Kent and Capello 2013), and the 2015 Nepal Earthquake (Poiani et al. 2016). VGI is also incorporated into citizen science projects including the Christmas Bird Count (Link et al. 2003), meteorological observations (Haklay 2013), and a wealth of other environmental applications (Brandeis et al. 2017).

Location-based social media (LBSM), a subset of VGI, is increasingly implemented in the study of social processes over space. Due to its popularity with the general public (Zickuhr 2013), LBSM has the potential to be used as a supplement, or replacement, for conventional datasets such as U.S. Census products (Lee et al. 2016; Navratil and Frank 2013). As a result of its social and place-based nature, LBSM has the potential to inform researchers on users' spatial preferences, habits, and day-to-day activities, not to mention relationships between places. From an emancipatory perspective, such data present opportunities to elicit patterns of inequality (Shelton et al. 2014) and draw attention to the needs of marginalized groups.

Yet, many questions remain on the nature of these datasets. Much attention has been devoted to the data quality, particularly veracity, of open mapping and citizen science VGI projects (Goodchild and Li 2012; Goodchild and Glennon 2010; Haklay 2010; Flanagin and Metzger 2008; Goodchild 2008), but questions of validity, particularly surrounding LBSM, have largely been neglected. What are demographics of contributors, and what are their motivations for contributing? How can these data be used to study spatial processes, and what do critical inquiries on spatial patterns tell us about contributions? More fundamentally, what do these data tell us? This dissertation seeks to answer these questions in the context of geography and GIScience.

Dissertation parts and organization

Organized in three article dissertation format, the first article (Chapter II) focuses on the production of LBSM and users' motivations for contributing LBSM. Through a web-administered survey of university students, it addresses questions about LBSM users' behaviors and perceptions on various platforms, noting differences in terms of gender, race, and academic standing. It explores users' motivations for contributing VGI, how users view privacy, and the role of place in location-enabled posts. In this article, the goal is not to demonstrate that LBSM is unbiased; rather, the results of the first article serve to inform the interpretations of the second and third articles. To properly understand the spatial patterns produced in LBSM, it is first necessary to understand users' demographics, perceptions, and motivations. The research questions addressed in this chapter are:

1. Do significant differences exist across gender, race, or academic standing groups in the use of LBSM?

2. Are there significant differences in the way these groups perceive LBSM? For instance, why do people choose (or not) to attach locational information to social media content? Are some groups more concerned about privacy than others, and do usage patterns appropriately reflect these concerns?

3. Most importantly, in the grander scheme of VGI research, what are the implications of users' responses for researchers seeking to utilize LBSM as a data source to study socio-spatial processes?

This first article (Chapter II) serves as the foundation for the other two articles (see Figure 1). The other two articles are quantitative case studies that focus on spatial processes of LBSM production but vary in terms of spatial scale, type of content production, and geographic subfield. Both make use of a conventional data source, the U.S. Census, for assessment purposes. Comparing Twitter data to traditional datasets is telling in that it reveals the characteristics of locations that users prefer, albeit within a particular context. However, as pointed out by Longley et al. (2015), these data sources measure different processes. Census data describes only where people reside while Twitter data reflects the mobile nature of users that may or may not be indicative of users' home locations. Nevertheless, significant and insignificant relationships are both informative from a data quality perspective. They equally help build toward a better understanding of the data.

The second article focuses on tweets contributed with precise location (i.e. exact latitude and longitude) at the county scale with census tracts as the unit of analysis. Carried out in Harris County, Texas (containing most of the Houston urban area), this study employs a little used facet of users' Twitter account information: their language preference. Various spatial and non-spatial regression models are used. The number of users with an account language other than English within each tract serves as a dependent variable. Population, population density, median income, median age, percent foreign born, percent white, and number of employees serve as independent variables. Besides its advancement of knowledge on LBSM, the second article also importantly contributes to the subfields of language, ethnic, and urban geography. Specifically, the second article (Chapter III) addresses the following research questions:

1. Can conventional variables – population, population density, median income, median age, percent foreign born, percent white, and number of employees – effectively explain the locations where people are using languages other than English on Twitter within Harris County, Texas?
2. To capture the potential effect of land use type on content production, are residential variables (e.g., the first six variables mentioned above) sufficient to explain variation, or are non-residential variables (e.g., number of employees) more effective?
3. How does LBSM inform us about the behaviors of users and aspects of place?

The third article (Chapter IV) focuses on tweets contributed with general location (e.g., a city, neighborhood, or region) at a larger regional scale, with municipalities as the unit of analysis. Carried out across U.S. Census-defined incorporated places in Texas and Louisiana, this study utilizes content in users' profile descriptions along with tweet text. Counts of users referencing #BlackLivesMatter or #AllLivesMatter in a tweet or their Twitter profile are used as dependent variables in four ordinary least squares (OLS) regression models. Population, population density, percent white, median income, median age, and percent unemployed from the U.S. Census are used as independent variables. The results of this third study have important implications for the geographies of race and ethnicity in addition to its contributions on the nature of LBSM data. Chapter IV addresses the following research questions:

1. Which census variables best predict the production of #BlackLivesMatter and counter-protest content in Texas' and Louisiana's cities?
2. Which cities are outliers, and what do tweets from individual users tell us about protest and counter-protest?
3. More generally, how can data from Twitter inform us about socio-spatial processes?

The results of the three articles are synthesized in Chapter V. In this chapter, I also include some future directions for research on big spatial data based on questions raised in this dissertation.

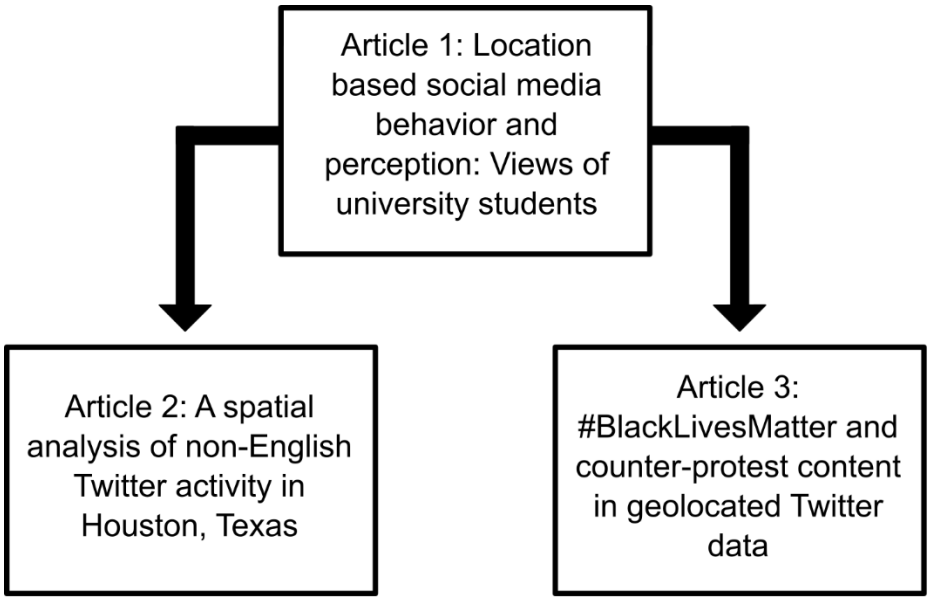

Figure 1.1 Structure of this three article dissertation

CHAPTER II

LOCATION-BASED SOCIAL MEDIA BEHAVIOR AND PERCEPTION: VIEWS OF UNIVERSITY STUDENTS[1]

Introduction

Connectivity in today's digital world

Over the previous two decades, society has witnessed a dramatic rise in the prevalence of mobile communications, Web 2.0 applications (O'Reilly 2005), web GIS, and cyberspace. Not only can people communicate over a free range of geographic space and participate in the web's social construction, these tasks can be performed simultaneously through the GeoWeb. While technological improvements have freed individuals from the constraints of static communication, people have not necessarily been 'liberated' from place (Malpas 2012); individuals are still part of and influenced by the relational networks of places. For many, much of today's "presentation of the self" (Goffman 1959, 2) occurs online, and Global Navigation Satellite Systems (GNSS; e.g., GPS, GLONASS, etc.)-enabled mobile devices and internet connectivity have allowed users to easily attach geographic information to web content (Goggin 2012), facilitating a social display of one's locational activities. Arguably, the role of location in communication and social networks is more important than ever before.

[1] Published as: Haffner, M., Mathews, A. J., Fekete, E., Finchum, G. A. (2017). Location-based social media behaviors and perception: Views of university students. *Geographical Review*, 108(2), 203-224.

Today, 88 percent of U.S. adults use the Internet, seventy-seven percent own a smartphone, and 69 percent use social media (Smith 2016). These figures have increased remarkably over the past ten years and are likely to continue increasing. Interestingly, 30 percent of these social media users have tagged their location in a post (Zickuhr 2013). This practice, termed location-based social media (LBSM), is a subset of volunteered geographic information (VGI) and has become the principal means by which people share locational information online (Goodchild 2007a, 2007b). With the proliferation of these activities, the use of LBSM as a data source for studying spatial processes has become common. In light of this, it is crucial to evaluate the meaning of this information and its validity for such purposes. Many studies have examined the demographics, perceptions, and motivations of those who contribute to explicitly geographic forms of VGI (for example, OpenStreetMap (OSM)), but there is a lack of research on implicitly geographic forms of VGI (such as Twitter and Instagram). Further, researchers have yet to determine the applicability of broad principles of VGI to LBSM.

Research focus

In this study we assess the demographics, usage patterns, and perceptions of a group with high rates of social media (Greenwood, Perrin, and Duggan 2016) and LBSM usage: university students (Zickuhr 2013). Through the administration of a Web-based survey, we address the following research questions: Do significant differences exist across gender, race, or academic standing groups in the use of LBSM? Are there significant differences in the way these groups perceive LBSM? For instance, why do people choose (or not) to attach locational information to social media content? Are some groups more concerned about privacy than others, and do usage patterns appropriately reflect these concerns? Most importantly, in the grander scheme of VGI research, what are the implications of users' responses for researchers seeking to utilize LBSM as a data source to study socio-spatial processes?

We discover that the LBSM user base is different from other VGI platforms and seemingly less biased. Females are the more common users and surprisingly are less concerned about privacy. More generally, place is an important social media component to a substantial number of users, making a compelling case for the use of such information in geography.

Background

Web 2.0 production

With the increasing prevalence of new forms of technology, "coded" processes—those hidden from users— have gained greater potential to alter interactions in physical space through covert power structures (Dodge and Kitchin 2005; Graham, Zook, and Boulton 2012). Authoritative representations on Web maps, such as those created by Google, are often accepted as objective, when realistically much is masked behind cryptic software (Zook and Graham 2007). The greatest online presence is available to those with the ability to purchase it, leading to uneven representations online and subsequently in physical space. Information agencies, therefore, act as gatekeepers, determining who is and who is not represented.

Neither the top-down representation of geographic information by companies nor the bottom-up production of GeoWeb content by users is evenly distributed. In the Web 2.0 era, the Internet is a socially formed space with many offline biases merely being reflected online. Kalev Leetaru and others (2013) and David Parr (2015) found that a minute number of contributors produced a disproportionately large amount of content. Researchers also noted a "digital divide"—unequal access to digital devices and connectivity—with much of this being expressed geographically (Warf 2001). Google Earth, which debuted prior to Hurricane Katrina and served as a medium for users to add their own placemarks, revealed noticeably fewer locations in the Lower-Ninth Ward of New Orleans, a predominantly low income, black neighborhood (Crutcher and Zook 2009). Similarly, there is an underrepresentation of geotagged Wikipedia articles in less-

developed countries (Graham and others 2014) and far fewer Foursquare locations in low-income, black neighborhoods compared to more affluent, white neighborhoods in U.S. cities (Fekete 2015). Despite more ubiquitous physical access to digital devices, a digital divide is still present. Some may not participate due to lack of Internet access or knowledge. Others may consciously choose not to utilize particular (or any) forms of social media. In this way, the manifestation of the digital divide is constantly changing (Crutcher and Zook 2009). Today, the greater divide is social, with biases being reflected on the Web. These persist despite claims that the Internet is, or will become, a purely democratic space (Warf 2001). Though the bottom-up approach of crowdsourcing and VGI has the potential to bring about more democratic forms of mapping and knowledge, it has yet to destabilize current authoritative forces (Kay, Zhao, and Sui 2015).

Web-based contributors

While Web maps are commonly used to study geographical biases of VGI, surveys are often used to study its social biases. Through this medium, differences have been uncovered with respect to gender, educational attainment, income, race, and age (see Bartoschek and Kebler 2013; Mathews and others 2013; Stephens 2013; Zickuhr 2013). Though the use of social media varies little with respect to gender, in general males are heavier contributors to VGI and some Web 2.0 platforms. Males geotag photographs more frequently (Stephens 2013), have greater awareness of OSM, and contribute more often to OSM than females (Haklay and Budhathoki 2010). Gender differences are expressed not only in rates of contribution, but through incentives for contribution as well. Thomas Bartoschek and Carsten Kebler reported that males were more likely to contribute to OSM with financial compensation, while females were more likely to contribute with “better usability” (that is, user friendliness) and if friends were active on the network (2013).

Bartoschek and Kebler also noted differences in rates of contribution based on educational attainment (2013). After class instruction with OSM, over 25 percent of university students were

regularly active contributors to the mapping project, as opposed to less than 10 percent of the high school students. This would suggest that higher educational attainment leads to greater rates of contribution. However, the findings of Adam Mathews and others were more mixed (2013). These authors reported that education, particularly geographic knowledge, played a positive role in risk awareness and location-based services (LBS) usage, but did not greatly influence privacy concerns or contribution likelihood.

In addition to educational attainment, income can also impact VGI participation and contribution. After analyzing the "intraurban divide" through restaurant reviews, James Baginski, Daniel Sui, and Edward Malecki (2014) did not observe a correlation between income and number of reviews. Instead, they noted a connection between menu price and reviews. They wrote, "the positive relationship between reviews and restaurant pricing support the idea that wealthier individuals dining at more expensive restaurants are more likely to contribute to Web 2.0 applications" (Baginski, Sui, and Malecki 2014, 449). While those with higher incomes are more likely to have access to Web 2.0 devices, this segment of the population may preferentially contribute restaurant reviews while ignoring other platforms. Currently, there is little evidence showing that people choose to contribute to all Web 2.0 (or VGI) platforms equally, or that those with higher incomes contribute to all Web 2.0 applications at a higher rate than those with lower incomes. In fact, Kathryn Zickuhr (2013) noted no significant difference in "geosocial service" usage (that is, location check-ins) due to income. In her survey, the highest income category utilized location check-ins the least.

These conflicting results unveil the difficulty of making generalizations across VGI platforms. One potential source of this discrepancy is in the varying nature of contributable geographic information, namely explicit versus implicit (Graham and Shelton 2013). On platforms such as OSM, Google Maps, and Wikimapia, information is explicitly geographic (that is, users work directly with Web maps), whereas on LBSM platforms such as Twitter, Facebook, and

Instagram, information is implicitly geographic (that is, users decide whether or not to attach location to content (Haffner and Mathews 2016). Further, within LBSM, even a single platform can exhibit significant variability. A Twitter post can be geotagged, simply mention a location, contain a photograph, or have any combination thereof. These cases exhibit varying degrees of spatiality which could drive unequal application by various groups (Haffner and Mathews 2016). Explicitly geographic platforms (for example, OSM) require greater place knowledge and greater effort in contribution, whereas implicit platforms, such as Twitter, merely require participation. Due to this, these latter platforms are more accessible to, and therefore potentially more representative of, marginalized populations. Indeed, Maeve Duggan found in the United States that Twitter, Tumblr, and Instagram were proportionally used by greater percentages of Hispanics and blacks than whites (2015).

Despite meaningful research on "racialized cyberscapes" (Crutcher and Zook 2009; Fekete 2015), the effect of race and ethnicity on contribution has been relatively unexplored with the exception of Zickuhr (2013) and Duggan (2015). Contrary to the assumptions that the elite are the most common users of the GeoWeb, Zickuhr (2013) found greater usage by minorities and women (2013). Those identifying as Hispanic, black, and white (exclusively) used geosocial services at rates of 24 percent, 11 percent, and 10 percent, respectively. In addition, females used geosocial services more frequently than males, though this difference was not statistically significant. Zickuhr further found that gender, educational attainment, and household income did not affect the rates at which social media accounts were locational-enabled (2013). Significant differences exist only for age, with greater rates for younger users. The effect of race was not examined with respect to location tagging.

People use LBSM for a variety of reasons. Some use it as a means of connecting with a particular place or community (Frith 2012; de Lange and de Waal 2013) by showing support for local businesses (Cramer, Rost, and Homquist 2011; Lindqvist and others 2011), participating in

political discussions (Gordon, Baldwin-Philippi, and Balestra 2013), and collectively solving complex urban problems (de Lange and de Waal 2013). Conversely, people used LBSM recreationally. Platforms like FourSquare double as location-based games and have become immensely popular (Farman 2012; Frith 2012). While connecting with friends is often an important part of such games and LBSM more generally (Frith 2012), personal satisfaction also drives LBSM usage (Goodchild 2007a; Humphreys 2013). On this note, Raz Schwartz and Germaine Halegoua (2015) have pointed out that location sharing can be merely another avenue of self-expression. In some cases, however, personal reasons for contribution should be classified as self- promotion (Goodchild 2007a; Evans 2015). In line with this, Leighton Evans contends that location sharing is used as a mechanism for building social capital (2015). Similarly, Matthew Wilson (2012) called location sharing through social media "conspicuous mobility," pointing out that people share their location when and where they want to be seen.

Platforms have changed markedly since Zickuhr's assessment and since initial efforts to utilize LBSM in geographic studies (2013). In Twitter's case, users previously had to enable location features deep within their account settings in order to geotag posts (Leetaru and others 2013). This made using location features relatively cumbersome and not intuitive. As of April 2015, the Twitter applications for Android (APKMirror 2015) and iOS (AMC 2015) have a location icon that appears on the prompt of every tweet (Twitter 2016). Along with this update, users are now able to specify location at a variety of scales (for example, a neighborhood, campus, city, or state), instead of being restricted to precise location in the form of latitude and longitude. While users still have the option to use precise location, most are using "general location" today, as confirmed in our survey. Other changes have enabled users to "push" posts to Twitter from other social media platforms, notably Foursquare and Instagram, and these posts also have the capacity to hold location information. These various changes, coupled with the growing popularity of using LBSM

in geographic research, warrant a more in-depth analysis of LBSM users' demographics and perceptions.

Data and methods

Data and collection procedure

To answer our research questions and address the identified research gaps, we administered a LBSM-focused questionnaire to university students. Data were captured using a Web-based survey created and hosted on Google Forms online. Due to university students' reputations as heavy users of social media and its locational convenience as a sample population, Oklahoma State University students, both undergraduate and graduate, made up the target population for the survey. In fall of 2015, Oklahoma State University had 25,806 total students of which 81.6 percent were undergraduates and 18.4 percent were graduates (IRIM 2015). The campus gender ratio is fairly even with slightly more males (51.7 percent). Most students are white (66.9 percent) with the remainder multiracial (8.1 percent), Hispanic (5.8 percent), African-American (4.5 percent), Native American (4.4 percent), Asian (1.7 percent), Pacific Islander (0.1 percent), and unknown (0.7 percent). Not included in these percentages are international students, who make up a sizeable portion of the student body (7.7 percent). The university has a particularly large Native American population, many of whom identify as Native American and another race.

Prior to email distribution of the survey, we conducted a pretest with a random selection of students on campus. Those willing to participate were asked to complete the form and reflect on it using a provided iPad. Student comments were then used to clarify ambiguous statements and identify possible gaps. Due to comments on Instagram's popularity, a section specific to this platform was added. Subsequently, a random subset of 5,000 university students (conforming to the bulk-email restrictions of Oklahoma State University) was emailed a message with a hyperlink to the survey. No compensation was offered for its completion. Consisting of forty-six questions,

the survey was organized into seven sections: (1) demographics, (2) mobile phone and tablet use, (3) social media use, (4) geotagging of social media, (5) Instagram use, (6) Twitter use, and (7) LBSM perceptions. All questions in the first six sections were multiple choice with the exception of "other" options with write-in text boxes and one question in the demographics section with a write-in for the student's academic major. Section 7 consisted of both open ended questions and five-point, Likert-scale perception questions ranging from "strongly disagree" to "strongly agree." The open-ended questions inquired about why students do or do not geotag their social media content. The Likert-scale questions pertained to privacy, negative consequences related to geotagging, and where students geotag (home versus away from home). The sections were embedded with logic to redirect respondents when appropriate. For example, if a respondent marked "No" on the question relating to (6) Twitter use, they were immediately redirected to the (7) LBSM perception section.

Data analysis

A total of 253 students completed the survey (slightly over 5 percent response rate). Overall, the sampled respondents were representative of the campus population (see Table 1). However, we generally had more females, whites, and graduate students participate. These participation levels are consistent with similar studies noting greater female (Sax, Gilmartin, and Bryant 2003; Mathews and others 2013; Stephens 2013) and graduate student (Mathews and others 2013) participation. We had low raw frequencies for most nonwhite groups, but the overall percentages did not drastically differ from those of the general student body. Yet, since all individual nonwhite groups had frequencies less than thirty, we did not separate each group in our analysis. Instead, we use two groups where nonminority includes only those identifying as white, and minority encompasses all others, including those identifying as both white and another race. We acknowledge that results with minority versus nonminority categories must be interpreted with caution, since experiential differences exist between various minority groups. However, we did not

want to exclude comparisons by race, particularly in light of recent discussions of race, geography, and Web 2.0 production (Crutcher and Zook 2009; Fekete 2015; Shelton, Poorthuis, and Zook 2015). Additionally, the nonwhite groups together accounted for a sizable portion of our respondents at 27 percent. As a measure of educational attainment, we use a collapsed version of our academic standing question, consolidating six categories into three: underclassmen (freshmen and sophomore), upperclassmen (junior and senior), and graduate students (master's and doctoral).

In our analysis, we implement a variety of statistical methods using the R Project for Statistical Computing (R Core Team 2016). In exploring the differences between groups in the use of LBSM, we construct m by n contingency tables on which we employ chi-squared tests. Due to the potential error induced in the two by two case, we use Yates' Continuity Correction. This correction subtracts 0.5 from every observed value in the table, resulting in more conservative tests. In our survey, we asked participants about the use of eight social media platforms—Twitter, Instagram, Google +, Facebook, Pinterest, Snapchat, Foursquare, and Flickr—and report the percentage use by each group. To avoid tests on exceedingly low cell frequencies, we test for significant differences in only the four most used platforms: Twitter, Instagram, Facebook, and Snapchat. We also explore differences beyond simple LBSM use, specifically investigating how users prefer to post (for example, cell phone, tablet, or desktop computer) and what kind of location (for example, general, precise, or both) Twitter users utilize. For the former, we again use a chi-squared test. For the latter, we use Fisher's Exact Test, which is analogous to the chi-squared test but effectively handles the presence of many low- (or zero-) value cell counts (Agresti 1990). Additionally, it is more conservative than chi-squared and computes exact p-values as opposed to approximations.

To gauge LBSM perception differences, we focus on Likert-scale responses to two statements: "Geotagging social media posts is a good way to let my friends and followers know where I am and what I am up to," and "I feel that geotagging infringes on my privacy." Here, we

employ two ordinal logistic regression (OLR) models. This technique explains an ordinal level dependent variable (such as Likert- scale items) and several independent variables on any scale, including nominal or ordinal, (such as gender, minority status, or academic standing). Similar to ordinary least squares (OLS) regression, the test produces a coefficient, standard error, and a t-value for each variable. In addition to these models, we qualitatively summarize the open-ended responses by gender to the questions "Why do you choose to geotag social media posts?" and "Why, at times, do you choose not to geotag social media posts?"

Evaluating whether or not students' perceptions align with their behavior is a difficult task. We intended to compare Likert-scale perception responses with frequency of geotagging on various platforms, but in general, students do not geotag many posts. For instance, of 118 students that geotag posts on Twitter, only eleven do so more than 25 percent of the time. Due to this, we instead compare the relationship between geotagging and the public/private nature of users' profiles with responses to statement "I feel that geotagging infringes on my privacy." To test for independence between groups and take into account the ordinal nature of this dependent variable, we use the Mann Whitney U test, a rank-based nonparametric method.

Results

Behavior: Differences in LBSM use and geotagging

Clear behavioral differences are apparent between groups in the usage of various social media platforms and geotagging on those platforms (Table 2). Not all are statistically significant, however. Dissimilarities exist for gender and academic standing, but not for race (Table 3). A greater number of females prefer to post social media content via mobile phone or tablet ($p = 0.011$), and more females use Instagram, Facebook, and Snapchat ($p = 0.000$ for all three; Twitter results were insignificant). Although a far greater number of females have geotagged some form of social media content ($p = 0.000$), more geotag by platform only for Facebook ($p = 0.048$). Comparisons across

academic standing groups show few notable differences between under- and upperclassmen in social media use and geotagging. These groups diverge only in Snapchat use, which is greater for underclassmen ($p = 0.007$). Many differences exist between undergraduate and graduate students. These are significant for Twitter, Instagram, and Snapchat use ($p = 0.046$, $p = 0.000$, and $p = 0.000$, respectively) and Instagram, Facebook, and Snapchat geotagging ($p = 0.003$, $p = 0.003$, and $p = 0.000$, respectively). Similarly, the differences between upperclassmen and graduate students are significant for Twitter, Instagram, and Snapchat use ($p = 0.014$, $p = 0.004$, and $p = 0.001$, respectively) as well as Facebook and Snapchat geotagging ($p = 0.013$ and $p = 0.001$, respectively). Graduate students deviate from both undergraduate groups in their preference of social media posting device ($p = 0.000$), viewing smartphones and tablets less favorably.

Aside from comparing geotagging use to the lack thereof, we also compare the use of Twitter's various location types—general location, precise location, both, neither, and unsure/default setting—across the three groups. The results displayed in Figure 1 include those who use Twitter but do not geotag posts on this platform, but the statistical tests (see Table 4) omit all those who do not geotag posts on Twitter. Here, the only noteworthy differences are between males and females ($p = 0.039$).

Perception

Prior to conducting the OLR analysis, females were coded with a one and males with a zero, minorities coded with a one and non-minorities with a zero, and graduate students coded with a two, upperclassmen coded with a one, and underclassmen coded with a zero. The first OLR model (Table 5) explains the statement "Geotagging social media posts is a good way to let my friends and followers know where I am and what I am up to." The only significant variable in this model is gender ($p = 0.007$), with females more in agreement with the statement comparatively (see Figure 2). The second model (Table 6) explains the statement "I feel that geotagging infringes on my

privacy." Despite apparent differences in percentages (Figure 3), no variables are significant in this model. The open-ended questions reveal subtle differences as to why students choose (or not) to geotag. Here, the greatest differences are between males and females. For why students choose to geotag, many terms are common to both males and females such as "people," "location," "show," "know," "place," and "post," but females list more terms, with many related to travel such as "vacation," "travelling," and "restaurant." Though the words "people" and "friends" show up for both males and females, "family" appears for females only. Regarding why students choose not to geotag, males responded with more terms comparatively. Again, some common strands are found in words such as "location," "people," "don't," "want," and "know." "Privacy," however, is more common for males, with words such as "advertiser," "never," and "expose" also present. Females lack these latter terms, but return the word "creepy." Individual responses are reviewed in greater detail in the discussion section.

Since the fundamental geotagging behaviors of each group are different, we could not confidently compare LBSM perception to behavior by group. Rather, we compare geotagging perception with geotagging behavior and metrics related to privacy—the use of public versus private profiles—across our entire sample. The Mann-Whitney U test (Table 7) reveals that agreement with the statement "I feel that geotagging infringes upon my privacy," is significantly different for those who have geotagged versus those who have not ($p = 0.000$) and for those who have a public versus private Instagram account ($p = 0.006$). Intuitively, those geotagging do not feel that it infringes upon their privacy, while those not geotagging do feel that it infringes upon their privacy. Similarly, those who have public Instagram accounts do not share privacy concerns over geotagging, but those with a private Instagram account do feel that geotagging infringes upon their privacy. On the other hand, public Twitter account holders do not express significant concern over their privacy when geotagging ($p = 0.149$), at least not to the same degree as those who have private Instagram accounts.

Discussion

Social media usage and geotagging on social media

The greatest differences in social media use and geotagging are between males and females, and between underclassmen and graduate students. Some of these results, such as greater activity from females, are consistent with the findings of other studies (see Stephens 2013). However, our findings on geotagging differ from those of Muki Haklay and Nama Budhathoki (2010), who find that an overwhelming majority of OSM contributors are male, and Monica Stephens (2013), who reports that more males geotag photographs. In our study, more females geotagged irrespective of social media outlet. The difference between our study and others highlights a curious phenomenon with several possible causes. Contribution to the GeoWeb by gender might be changing. While high income, highly educated white males may have been early adopters, over time accessibility has increased for other segments of the population. Alternatively, the implicitly geographic nature of LBSM might be intrinsically more attractive to others. This result, differing from those in studies of explicit forms of geographic information, suggest that principles of one VGI platform cannot be wholly extended to others. Even within LBSM, principles of contribution cannot be applied across platforms. Supporting this, 55 percent of our respondents either agree or strongly agree with the statement "I am more likely to geotag on certain social media platforms (for example, Twitter only) and not all of the social media platforms that I use" while only 20 percent either disagreed or strongly disagreed.

While statistically significant, the dissimilarity between males and females in the types of location used on Twitter is somewhat deceiving. The largest difference between the two groups are of those falling into the "Unsure/default setting" category. If users possessed a current version of the Twitter application and were using the default setting at the time of the study, they were unknowingly using "General location (for example, city, neighborhood)." Adjusting for this

discrepancy results in comparable use of Twitter's location types between males and females, and also within race and academic standing.

Perception

Echoing the LBSM behavior findings, gender is the greatest driver of LBSM perception. Females view geotagging more positively and surprisingly are less concerned about privacy. Given that more females geotag and use social media, it is presumable they would only do so if lacking major privacy concerns. An evaluation of individual responses to the question "Why do you choose to geotag social media posts?" sheds light on this. While numerous male and female students state that they attach location to posts to demonstrate their location to friends, a telling portion communicate that location is simply a logical extension of the post itself. The following responses illustrate this clearly:

- "Usually it's just to further emphasize the caption."
- "Because my location is part of the story being told by my post."
- "It adds additional information to my post and/or it gives a better understanding of what my post means (such as when on vacation)"
- "[B]ecause where I am is important to the post."

Seeing how students view location in these responses, it is likely that place is a common topic in posts regardless of whether or not the post is geotagged. Thus, it is also probable that if students do not have privacy concerns over posting social media content in general, they do not have concerns about tagging their location either. Further, many students also claim to not geotag posts primarily if location is not relevant, neglecting to mention concerns over privacy or explicit wishes to conceal their location. Interestingly, six females (but no males) mention that they use

LBSM to keep a personal record of places visited. This way, geotags function as a kind of locational journal:

- "So that I can look back and remember where I was and what I was doing at certain times."
- "To remember where I was at the time of the post."
- "So I can remember where and what I did in my life."

While these responses do not make privacy concerns void, these females make clear that they are posting for themselves and not for others.

An examination of individual responses to the statement "Why, at times, do you choose to not geotag social media?" confirms the Likert-scale responses about privacy with respect to gender. Only 11 percent of females report concerns about privacy or safety as opposed to 28 percent of males. In this case, females may be less concerned about privacy because more females have private Twitter (35 percent) and Instagram (68 percent) accounts than males (18 percent and 50 percent, respectively), so fewer strangers see their posts anyway. Alternatively, while more females geotag posts, the way that they geotag may be different from males. For example, one female rationalizes:

- "To be honest, I hardly geotag unless the place is really cool and popular. Other than that, I find it strange to continuously post about my location. I feel as if people would creep on me, strangers of course, and that's not really pleasing to hear."

While impossible to ascertain without a more in-depth approach, it may be that females protect themselves by posting selectively in places where they feel safe, or by posting a nonintrusive general location. In terms of privacy and behavior across racial groups, minorities feel that geotagging infringes on their privacy (at least more so than others), but their social media and LBSM usage are not significantly different from nonminorities. Again, a possible explanation is

that minorities geotag but with private accounts so that only their friends and followers see their location. We find that a greater percentage of minorities (83 percent) have private Instagram accounts compared to nonminorities (56 percent), but more nonminorities (32 percent) have private Twitter accounts (32 percent compared to 19 percent for minorities).

Those who feel that geotagging infringes on their privacy are far less likely to geotag across all groups, but privacy with geotagging does not necessarily translate to profile privacy on all social media platforms. In general, those with private Instagram accounts feel that geotagging infringes on their privacy. This is similarly exhibited on Twitter, albeit to a much lesser degree. Of those who use Instagram and Twitter, far more have private Instagram accounts (63 percent) than private Twitter accounts (29 percent). This finding supports the view of Twitter as an "open forum" where users are more apt to share content with strangers. Interestingly, several newer LBSM platforms, such as Yik Yak and Tinder, require users to enable location and emphasize interaction with strangers. While our survey did not inquire about such outlets, future work may explore how users view privacy on these more anonymous outlets in addition to which users are concerned about disclosing their location with these outlets. In our study, several students indicate concern about strangers knowing their location, but what about certain friends, acquaintances, corporations, and/or the government? Such questions remain unanswered. The intricacies between social media use, geotagging, and the public/private nature of social media accounts are likely indicative of complex relationships that cannot be explained through surveys alone.

While these questions are beyond the scope of this paper, future research might benefit from a more in-depth qualitative approach implementing in-person interviews and/or focus groups to interpret these relationships. In addition to not inquiring about more anonymous LBSM platforms, this study has other limitations. While the overall sample size (253 respondents) was robust, low frequencies among several individual groups prohibited detailed analyses of race. We acknowledge that discrete "minority" and "nonminority" categories are not ideal, and the use of

these categories reveals more similarities than differences. A larger sample within individual groups would allow for an intersectional approach to compare the combined effects of gender, race, and academic standing. It is difficult to generalize across a category, such as gender, when the practices, experiences, and perceptions of nonwhite and white females may be vastly different. Additionally, our focuses on binary usage of LBSM rather than amount of use. A follow-up study could reveal demographic differences in the frequency of LBSM production. Although the sample in this study is representative of the university, it is not a suitable representation of the country as a whole (e.g., age and income biases on a college campus). Nevertheless, the bulk of our findings on usage rates align with those of Zickuhr (2013), whose sample consists of a wide range of education levels, ethnicities, and income groups across the entire United States.

Implications for researchers

The results of this study further challenge the notion that VGI is predominately contributed by the elite: white, high-income, highly educated males (Haklay and Budhathoki 2010; Bartoschek and Kebler 2013; Stephens 2013; Baginski, Sui, and Malecki 2014). Our sample reveals that females are more engaged with social media and LBSM, and minorities contribute at a rate comparable to nonminorities. While the footprint of the elite is certainly strong on explicitly geographic VGI platforms, such as OSM, the bulk of contributors to implicitly geographic forms of VGI, such as Twitter, are different. Some platforms are becoming more accessible to, and therefore more representative of, the general population. With regard to concerns that LBSM platforms are not representative of their user base, our study demonstrates that geotagging of social media does align with the demographic profile of each platform's users. In other words, LBSM may be no more biased demographically than social media in general. For researchers using LBSM in the study of spatial processes, this is clearly a positive finding. LBSM can potentially be used to reveal patterns of social disparity and represent the voice of marginalized populations.

Sophia Huyer and Nancy J. Hafkin claim that increasing females' confidence and dexterity with information and communication technologies will play an important part in bringing about gender equality (2006). In this context, the findings of greater female social media use and geotagging rates are encouraging. On the other hand, this may simply be a new manifestation of old processes. Though many groups now have physical access to the Internet, Richard Joiner, Caroline Stewart, and Chelsey Beaney contend that an often overlooked second digital divide persists (2015). This discrepancy lies in the reasons and attitudes toward use; such differences are plainly evident in our study. While computers and the Internet were initially designed by and for men, this practice excluded a large, untapped customer base (Cooper 2006). Today, social media is geared toward all, likely for monetary reasons. We should certainly remain critical of technological determinism when equality is motivated by profit.

On other grounds the GeoWeb persists as a nondemocratic space. Presence on LBSM does not guarantee an audience (boyd 2010), and connectedness on the Web may not be nearly as uniformly distributed as its usage. Additionally, those who produce content want to be noticed and do so where they want to be noticed (Wilson 2012; Evans 2015). Thus, LBSM may not be representative of users' everyday lives. Contributors have many motivations for posting content. In our study, users are driven by personal satisfaction, self-promotion, to connect with friends, and to promote businesses. The vast majority of students claim to post only at interesting and memorable locations, such as vacation spots. Since users can be selective with posting in this way, it is difficult to determine users' characteristics based solely on the location where content is produced. Using census data to determine users' characteristics is a promising prospect for VGI research, but an individual's place of residence cannot be inferred from one post alone (Shelton, Poorthuis, and Zook 2015). Indeed, in our study 77 percent of students either disagree or strongly disagree with the statement "I am likely to geotag posts at home (dorm, apartment, house, etc.)."

Aside from this, the fact that few people use precise location on platforms like Twitter is problematic for researchers. While general location more effectively preserves users' privacy, precise location is preferable to researchers because it is exact, limited in accuracy only by the capabilities of the device used, and allows for aggregation at a variety of spatial scales. Aggregation is difficult with general location, which is defined by large, generic polygons that do not align with standardized boundaries (as examples: census tracts, counties). Whereas Twitter users could originally only geotag posts with precise latitude and longitude, today the use of general location is increasingly common. In our sample, only two students report exclusively using precise location on Twitter, and only seven use both precise location and general location. It must be kept in mind that only a subset of the general population uses Twitter, a small percentage of tweets are geotagged, and very few users contribute precise location. This combination results in an extremely small percentage of the general population driving the production of precisely located Twitter data.

Conclusion

This study demonstrates several notable findings. Females and underclassmen are the most likely users of social media and LBSM, and few statistically significant racial differences exist in our survey. Facebook and Instagram are the most popular platforms for tagging location content for all groups, and overall students' behaviors align with their perceptions of LBSM. The greatest perception differences exist between males and females, and surprisingly, males are more concerned about privacy. This may be due to the fact that more females have private accounts to maintain greater control over those viewing posts, or they may post selectively in ways that ensure their safety. While this makes sense on a surface level, it is quite possible that something deeper is driving female enthusiasm and male skepticism toward geotagging.

Overall, LBSM users are representative of each social media platform's user base. LBSM usage, per our survey, is more representative of the general population than other forms of VGI.

For those wanting to study spatial processes with LBSM, these are encouraging findings. Student motivations should provoke a cautionary outlook though. They are motivated by a variety of factors including self-promotion, and many students alluded to only geotagging at unique or interesting locations. Nevertheless, it is clear that for students who do geotag, locational content is important. Many view geotagging as simply logical extensions to a posts that are likely already place-focused, and some (all females) use LBSM as a locational journal to keep track of noteworthy places visited. This finding aligns with Frith's notion that locative media influences how place is both perceived and experienced (2012).

Findings suggest that explicit and implicit forms of VGI be conceptualized differently. These two classes of platforms have different user demographics, motivations, and perceptions driving their production. LBSM's relative ease of use has somewhat leveled the playing field; many users without the means to contribute explicitly geographic VGI content can easily contribute to LBSM. While this may make LBSM a place of greater democratic opportunity, biases still exist, and privacy concerns abound. Despite these limitations, we remain in agreement with Mark Graham and Taylor Shelton (2013) and Harvey Miller and Michael Goodchild (2015)—that is, cautiously optimistic of the future of LBSM and its potential in geography.

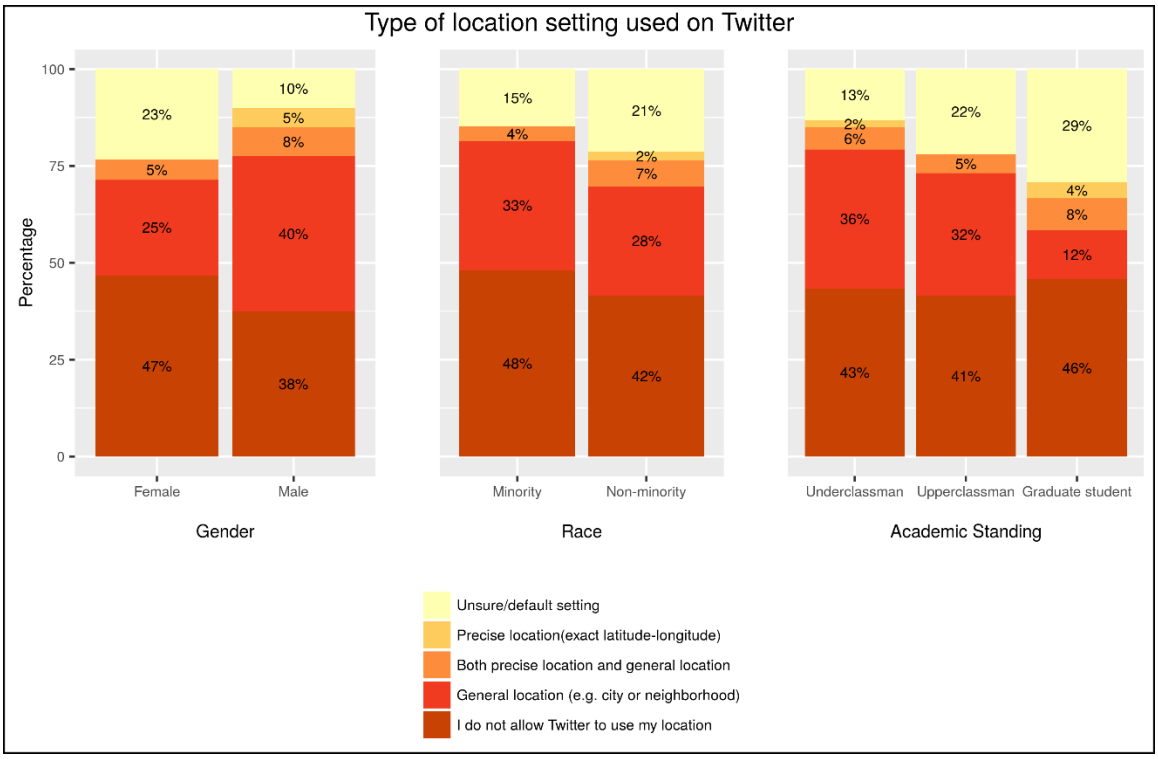

Figure 2.1 Type of location setting used on Twitter

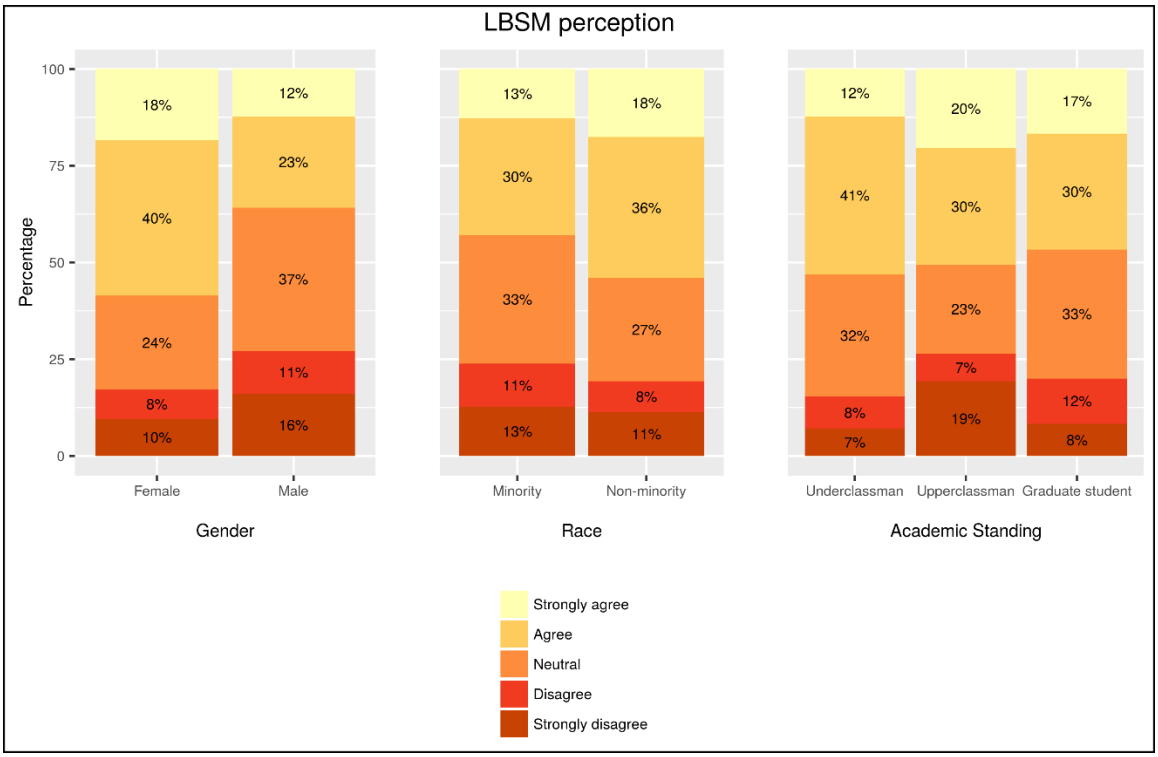

Figure 2.2 Response to the statement, "Geotagging social media posts is a good way to let my friends and followers know where I am and what I am up to".

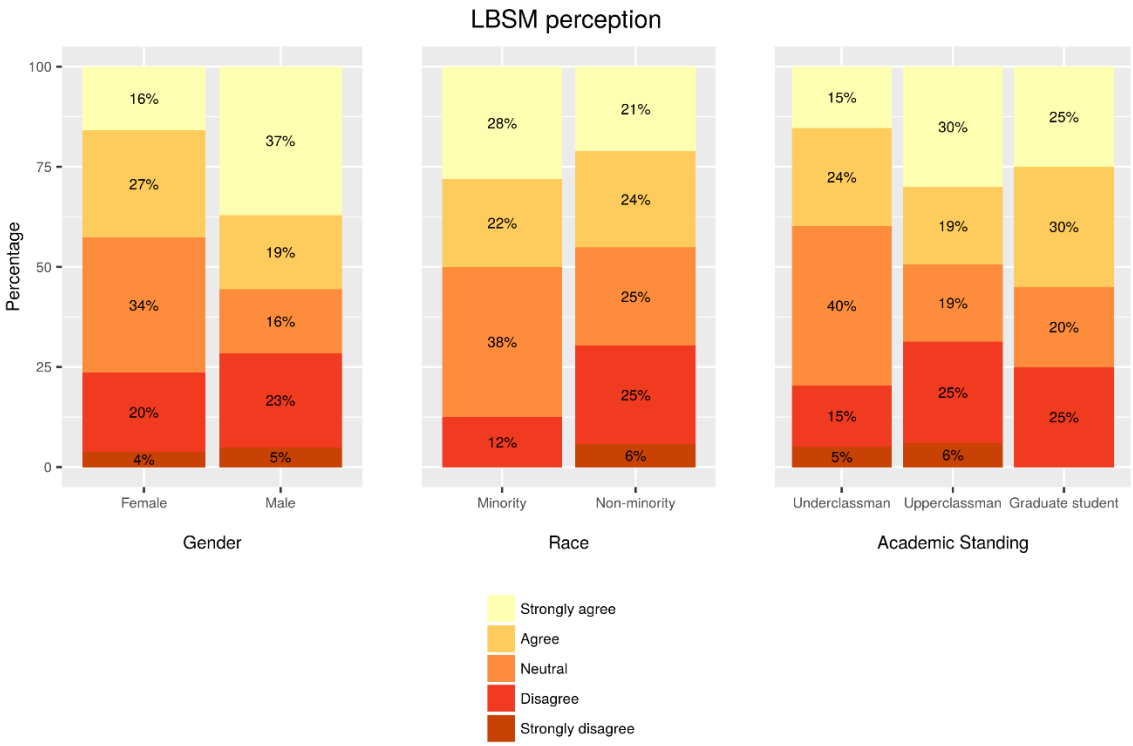

Figure 2.3 Response to the statement, "I feel that geotagging infringes on my privacy".

Table 2.1 Demographics of survey respondents

Category	Item	University %	Sample % (253)
Race/ethnicity	White	72.5%	73.0% (184)
	African American	4.9%	3.6% (9)
	Native American	4.8%	4.4% (11)
	Hispanic	6.3%	1.6% (4)
	Asian	1.9%	4.8% (12)
	Pacific Islander	0.1%	0.4% (1)
	Multiracial	8.8%	10.3% (26)
Minority status	Minority	26.7%	27.0% (160)
	Non-minority	73.3%	73.0% (88)
Academic Standing	Freshman	19.2%	20.2% (51)
	Sophomore	17.1%	19.5% (49)
	Junior	20.4%	17.1% (43)
	Senior	23.6%	16.7% (42)
	Graduate student	18.2%	26.2% (66)
Gender	Female	48.3%	34.9% (88)
	Male	51.7%	63.9% (160)

Table 2.2 Use of social media and LBSM

	Gender		Race		Academic Standing		
	Female	Male	Minority	Non-minority	Underclassmen	Upperclassmen	Graduate students
Use Twitter	55.6%	54.6%	51.5%	56.0%	58.0%	62.4%	40.9%
Use Instagram	80.0%	53.4%	67.7%	71.2%	83.0%	73.0%	48.5%
Use Google+	14.4%	10.2%	17.7%	10.9%	9.0%	10.6%	21.2%
Use Facebook	94.4%	77.3%	85.3%	89.1%	89.0%	91.0%	85.0%
Use Pinterest	68.1%	11.4%	29.4%	54.4%	45.0%	57.6%	39.4%
Use Snapchat	75.0%	51.4%	60.3%	68.5%	85.0%	67.1%	37.8%
Use Foursquare	0.6%	0.0%	0.0%	0.5%	0.0%	0.0%	1.5%
Use Flickr	0.6%	0.0%	0.0%	0.5%	0.0%	0.0%	1.5%
Have geotagged social media content	76.7%	51.9%	65.2%	69.5%	71.7%	67.9%	63.3%
Prefer to post social media content via smartphone or tablet	86.2%	71.6%	82.8%	80.8%	86.9%	88.1%	61.7%
Geotag on Twitter	8.8%	9.1%	10.3%	8.2%	11.0%	8.2%	6.1%
Geotag on Instagram	55.7%	29.6%	44.1%	46.2%	56.0%	47.1%	28.8%
Geotag on Google+	1.3%	0.0%	1.5%	0.5%	1.0%	0.0%	1.5%
Geotag on Facebook	58.1%	27.3%	47.1%	47.3%	45.0%	45.9%	53.0%
Geotag on Pinterest	1.3%	0.0%	1.5%	0.5%	2.0%	0.0%	0.0%
Geotag on Snapchat	31.3%	14.8%	23.5%	26.9%	37.0%	28.2%	4.5%
Geotag on Foursquare	0.6%	0.0%	0.0%	0.5%	0.0%	0.0%	1.5%
Geotag on Flickr	0.6%	0.0%	0.0%	0.5%	0.0%	0.0%	1.5%

Table 2.3 Chi-square test results for gender and race with social media/LBSM usage

	Gender		Race		Academic Standing					
					Underclassmen – upperclassmen		Underclassmen – graduate students		Upperclassmen – graduate students	
	X^2	*p*-value	X^2	*p*-value	X^2	*p*-value	X^2	*p*-value	X^2	*p*-value
Use Twitter	0.00	0.976	0.34	0.562	0.65	0.204	3.99	0.046*	6.02	0.014
Use Instagram	18.68	0.000***	0.02	0.888	2.18	0.140	20.66	0.000***	8.44	0.004**
Use Facebook	14.47	0.000***	0.17	0.676	0.01	0.911	0.30	0.583	0.68	0.409
Use Snapchat	13.47	0.000***	0.41	0.520	7.31	0.007**	37.41	0.000***	11.60	0.001**
Have geotagged social media content	14.22	0.000***	0.17	0.679	0.16	0.685	0.86	0.354	0.15	0.699
Prefer to post social media content via smartphone or tablet	6.64	0.011*	0.03	0.866	0.00	0.980	12.15	0.000***	12.38	0.000***
Geotag on Twitter[a]	1.08	0.299	0.21	0.650	0.09	0.767	0.21	0.650	0.00	1.000
Geotag on Instagram	0.92	0.337	0.00	0.985	1.15	0.283	9.08	0.003**	3.13	0.077
Geotag on Facebook	3.91	0.048*	0.01	0.911	0.75	0.091	8.53	0.003**	6.12	0.013*
Geotag on Snapchat	0.76	0.384	0.00	1.000	1.06	0.304	19.47	0.000***	11.49	0.001**

[a]Low, (but non-zero) frequencies; results must be interpreted with caution
* Significant at p = 0.05
** Significant at p = 0.01
*** Significant at p = 0.001

Table 2.4 Results of Fisher's test for type of location used on Twitter

	p-value
Gender	0.039
Race	0.803
Academic Standing	
Underclassmen – Upperclassmen	0.166
Underclassmen – Graduate Students	0.671
Upperclassmen – Graduate Students	0.063

Table 2.5 Ordinal logistic regression results for select variables predicting "Geotagging social media posts is a good way to let my friends and followers know where I am and what I am up to"

Variable	Coefficient	Std. error	*t*-value	*p*-value
Gender	0.69	0.25	2.70	0.007
Race	-0.30	0.26	-1.13	0.258
Academic Standing	-0.04	0.15	-0.25	0.801

Table 2.6 Ordinal logistic regression results for select variables predicting "I feel that geotagging posts infringes upon my privacy"

Variable	Coefficient	Std. error	*t*-value	*p*-value
Gender	-0.40	0.26	-1.54	0.124
Race	0.45	0.26	1.75	0.079
Academic Standing	0.13	0.15	0.88	0.377

Table 2.7 Mann Whitney U test results of 'I feel that geotagging infringes on my privacy' compared with LBSM behavior

Variable	*z*-score	*p*-value
Have vs. have not geotagged social media content	7.10	0.000
Public vs. private Instagram account	-2.74	0.006
Public vs. private Twitter account	-1.44	0.149

CHAPTER III

A SPATIAL ANALYSIS OF NON-ENGLISH TWITTER ACTIVITY IN HOUSTON, TEXAS[2]

Introduction

The integration of social media data into geographic research has become common (Sui & Goodchild, 2011), yet the question of social media's validity in such contexts is often overlooked. Social media data commonly suffers from demographic (Stephens, 2013; Zickuhr, 2013; Greenwood, Perrin, & Duggan, 2016; Haffner, Mathews, Fekete, & Finchum, 2017) and spatial biases (Crutcher & Zook, 2009; Leetaru, Wang, Padmanabhan, & Shook, 2013; Hecht & Stephens, 2014; Fekete, 2015) as well as disproportionate production of content (Elwood, Goodchild, & Sui, 2013; Shelton, Poorthuis, & Zook, 2015). Nevertheless, cities are becoming ever more reliant on such data for decision making (Kitchin, 2013), and these datasets can be effective if their limitations are properly understood (Miller & Goodchild, 2015). User-generated content, including social media, can meaningfully supplement conventional data sources, such as those from the U.S. Census (Goodchild, 2008; See et al., 2016), and bring about new ways of obtaining geographic knowledge (Miller & Goodchild, 2015). However, a more fundamental understanding of these datasets is required if they are to be used appropriately.

[2] Accepted for publication 25 February 2018 as: Haffner, M. (2018). A spatial analysis of non-English Twitter activity in Houston, Texas. *Transactions in GIS*.

Long have cities been hotspots for technological innovation, cultural exchange, and diverse communities. The production of social media data in cities – and the use of social media data by cities – is no exception. In the 21st century, local governments, urban planners, and city officials face unique challenges in meeting the needs of many groups, particularly the most vulnerable: ethnic and religious minorities, women, and children. Social media data present an opportunity to represent these groups in new ways, potentially leading to effective policy formation, but only if the data are shown *valid* for such purposes. In other words, the data are only useful if the assumptions about what they measure are correct. In this article, I attempt to gain a better understanding of what location based social media (LBSM) data actually represent by exploring its relationship with U.S. Census data. Using a subset of social media data potentially representative of ethnic minorities – tweets produced by users with an account language other than English – I use various regression techniques to evaluate the strength of relationships and draw special attention to outliers. I examine the influence of outliers in such scenarios – which is quite strong in this study – and discuss broader challenges for geographic research utilizing social media data.

Twitter, language, VGI, and conventional data

Geolocated Twitter data is a type of location-based social media (LBSM), falling under the broader umbrella of contributed geographic information (CGI) (Harvey, 2013). LBSM is different from common forms of volunteered geographic information (VGI) (Goodchild, 2007) in that the spatial information is often implicit rather than explicit (Graham & Shelton, 2013). With implicit spatial information, alternatively termed ambient geospatial information, the spatial aspects are secondary to other characteristics of the data (Stefanidis, Crooks, & Radzikowski, 2013). The implicit nature of spatial information in LBSM provides the opportunity to reveal more about users' everyday lives than more explicitly geographic forms of VGI, such as Open Street Map (OSM).

When a Twitter user attaches location to a post, often termed "geotagging," they have

multiple options. They can use general location (e.g., a neighborhood, city, or state), which is sent to the Twitter public application programming interface (API) in the form of four latitude-longitude pairs, creating a bounding box around their location. Alternately, users can choose precise location, which is displayed as one latitude-longitude pair, representing a point in the Twitter public API. This second form is much less common (Haffner et al., 2017), but its greater precision provides researchers with more flexibility (i.e. the ability to spatially aggregate and link to other spatial datasets). Patterns can be investigated at a variety of spatial scales.

In general, the study of language patterns through CGI is a sparsely researched topic. Leetaru et al. (2013), in a global analysis of geolocated Twitter activity, report that most geotagged tweets are written in English (41.57%), followed by Spanish (11.16%) and Portuguese (9.50%). Examining spatial patterns of language use, they find that Twitter reflects some expected patterns. In general, the languages used within European countries are reflective of the preeminent language in each place, and the effects of French and English colonization can be seen in tweets throughout Africa (Leetaru et al., 2013). Patterns less reflective of dominant languages exist in countries such as the Czech Republic, Austria, and the Balkan states where a wide variety of languages is exhibited on Twitter, perhaps more than the number of languages used offline.

At a finer spatial scale, Graham and Zook (2013) explore spatial patterns of language use by examining within-country geolinguistic contours of user-generated content on Google Maps. Comparing the prevalence of terms in competing languages (e.g., French versus English in Canada), they find that offline power relations are largely present online. In a similar fashion, Cheshire, Barratt, Manley and O'Brien (2016) and Cheshire, Manley, and O'Brien (2016) have produced maps highlighting the locations of tweets, symbolized by language, in both New York City and London. These maps show the top ten languages present in each city on Twitter, and illustrate recognizable spatial patterns. This project clearly exhibits a fascinating visual application of big spatial data. More generally, the works mentioned previously demonstrate the utility of using CGI data on language, supporting such avenues of research. However,

Crampton et al. (2013) suggest researchers think 'beyond the geotag' in academic contexts, advocating for more than the mapping of static points.

Combining user-generated sources with conventional data, such as various products from the U.S. Census, is one suggested way to do this. Longley, Adnan, and Lansley (2015) take this approach in comparing U.K. Census data with the gender, age, and ethnicity of Twitter users as determined through the Onomap classification. Li, Goodchild, and Xu (2013) compare the locations of Twitter and Flickr content to U.S. census data, finding that education has a positive effect on content production. Similarly, Kent and Capello (2013) compare demographic characteristics to the prevalence of geolocated Twitter data referencing the 2012 Horsethief Canyon Fire in Jackson, Wyoming. Using geographically weighted regression (GWR) they find that census blocks with a high percentage of population under 18 is the best predictor of content production. Griffin and Jiao (2015) use data from the cycling application Strava to find roads most frequently traversed by cyclists, and how these patterns correspond to the variables relating to the built environment. They suggest that planners use such results to find the most beneficial locations for bicycle lanes. Additionally, algorithms employing Twitter data have been proposed to predict travel demand using Twitter data in Los Angeles (Lee, Gao, & Goulias, 2016). Lee et al. (2016) suggest that such methods could serve as reasonable, up-to-date alternatives to household survey data for origin-destination trip estimation.

Studies using LBSM data often exhibit tension between providing new ways of acquiring geographic knowledge yet being fraught with problems. After eliminating erroneous observations, big datasets often become rather small (n = 217 in Kent and Capello's (2013) case). The "long-tail effect" – the consequences of a small number of users producing a disproportionately large amount of content (Elwood, Goodchild, & Sui, 2013) – heavily skews datasets. Further, all social media platforms incur some degree of demographic bias, and most lack capabilities to determine characteristics of individual users (Miller & Goodchild, 2015). Open Street Map is mostly used and contributed to by men, thus reflecting male local knowledge (Stephens, 2013). Similarly, Strava is a heavily male-dominated platform (Griffin

& Jiao, 2015), so any planning decisions made with Strava will reflect the preferences of this group. Determining users' home locations, work locations, and trip purposes is difficult, complicating the application of Twitter-based travel-demand modeling. Beyond these concerns, precise geotagging on Twitter appears to have dropped off considerably; in 2013, while only 2% of tweets were geotagged, over half of these contained precise location (Leetaru et al., 2013). A more recent survey of university students shows that only a small percentage of users who geotag tweets enable precise location (Haffner et al., 2017).

Kent and Capello's (2013) finding of a correlation between Twitter content and census blocks with younger populations would seem to suggest that younger users have a greater propensity to use LBSM to discuss hazards and disasters. However, it cannot be assumed that younger users are more likely to post about all disaster situations, and the residential characteristics of a location cannot be wholly applied to the people using social media in that location. The popularity of Foursquare (Fekete, 2015) and Facebook Places (Wilson, 2012) indicates that users very much enjoy posting at locations away from home, at least sometimes. Yet, this finding aligns with others confirming that younger individuals are the most likely to use social media (Zickuhr, 2013; Greenwood et al., 2016). Similarly, other research has confirmed offline biases being manifested on the GeoWeb. Foursquare contains far fewer locations in poor, black neighborhoods, reflective of the lack in opportunities in physical space (Fekete, 2015). Shortly after Hurricane Katrina the newly launched Google Earth showed a lack of placemarks in the Lower Ninth Ward, a low-income, black neighborhood (Crutcher & Zook, 2009). From a data validity perspective, these results are encouraging, but we must be cautious of spurious patterns (Miller & Goodchild, 2015; Zook et al., 2017).

Research objectives

Studies using LBSM demonstrate its potential in addressing urban problems, but many questions remain unanswered on its nature. Is LBSM representative of individuals' day-to-day travel patterns? How well does LBSM correspond with conventional data sources? Kennedy

(2006) has demonstrated that online self-expression is largely consistent with offline attitudes, so important parts of identity, such as language, race, and ethnicity, should manifest themselves online as well. While an individual's ethnicity is not directly discernible through social media, language use is more measurable. In fact, Twitter's streaming API reports language data in two ways: through each tweet's text using a language detection algorithm and through users' account settings. Conventional data sources (e.g., U.S. Census), collect data on these three characteristics, most notably on race and ethnicity. Since language, ethnicity, and race are often closely connected (Trimble & Dickson, 2007), albeit with some exceptions (e.g. Hispanics who use English exclusively but do not speak Spanish, and white Europeans who do not speak English), one would expect a certain degree of correspondence between this new online data source (Twitter) and conventional data sources (U.S. Census). That said, any data source is inherently limited in scope and only valid within its intended domains. Many U.S. Census products are limited to residential or employment characteristics of places, which, undoubtedly measure different phenomena than LBSM data. For this reason, the analysis of relationships between data sources in this study is exploratory, with the goal of gaining a better understanding of LBSM.

This study utilizes a spatial analysis of Twitter activity in Harris County, Texas (the heart of the Houston Metropolitan Area), focusing on users with an account language other than English. The approach is "abductive" in nature (Miller & Goodchild, 2015), focusing on unique relationships between users and places, leading to the formation of hypotheses about the data. Specifically, I ask the following research questions: (1) Can conventional variables – population, population density, median income, median age, percent foreign born, percent white, and number of employees – effectively explain the locations where people are using languages other than English on Twitter within Harris County, Texas? (2) To capture the potential effect of land use type on content production, are residential variables (e.g., the first six variables mentioned above) sufficient to explain variation, or are non-residential variables (e.g., number of

employees) more effective? (3) Finally, how does LBSM inform us about the behaviors of users and aspects of place?

Data and methods

In addressing these questions, several forms of regression are used: ordinary least squares (OLS), spatial autoregressive models (SAR; e.g., spatial lag and spatial error models), and geographically weighted regression (GWR). The spatial lag model treats a lagged version of the dependent variable as a new independent variable, whereas the spatial error model attempts to compensate for spatial dependence in the errors (Anselin, 1988). GWR, on the other hand, computes a local regression at each observation using values from other nearby observations (Brunsdon, Fotheringham, & Charlton, 1996). Upon finding deficiencies in each model, subsequent models are examined in the order described above. Following GWR, non-stationary variables are examined in greater detail. These models are all used in an exploratory fashion, not as a means of prediction.

The dependent variable (DV) is calculated from precisely geotagged tweets (i.e. those with exact latitude-longitude coordinates) and represents the total number of Twitter users with an account language other than English within each census tract. For brevity, this variable is referred to as non-English Twitter Users (NETU). The data were collected using Twitter's streaming application programming interface (API) and the Python module 'Tweepy' from 17 October 2015 to 26 November 2016. Census tracts (n = 786) were chosen as the unit of analysis in an effort to extract maximum detail while avoid sparseness. Census tracts are small enough to show significant variability with many variables, and most census tracts contain at least one tweet from a NETU. Blocks and block groups, on the other hand, are much sparser.

Counties in Texas were initially evaluated as candidate study areas due to the presence of several large, international population centers (e.g., Houston, Dallas, and San Antonio) and a wealth of native and foreign born people who speak languages other than English, most notably Spanish. Upon investigation, Harris County (see Fig. 1) possesses the greatest number

of tweets contributed by NETU, accounting for roughly 24% of all Twitter activity from this subset of users in Texas. Harris County overlaps with the Houston metropolitan area and has by far the largest population of any county in the state. According to the Migration Policy Institute, Houston is the most diverse of the country's ten largest metropolitan areas (Capps, Fix, & Nwosu, 2015). Most of its foreign born population is from Mexico, accounting for the 45% of the area's immigrants, followed by those from El Salvador, Vietnam, India, and Honduras, respectively. Interestingly, no racial or ethnic group is a majority, with non-Hispanic whites making up 40% of the population, followed by Latinos (36%) and blacks (17%).

When a user creates a Twitter account, a profile language is determined automatically, defaulting to the language used to access Twitter during account creation. Regardless of account language, tweets are always displayed in their language of authorship (Twitter, 2017), but the account language determines other content, such the language of account settings, emails from Twitter, and notifications. Therefore, it is likely that a user's account language is one that the user understands, if not prefers. It is unlikely that a user would select an account language that they do not know since this setting is not visible to other users and would potentially impair their ability to use the platform. Twitter also determines the language of each individual tweet through a language detection algorithm, but this facet is not the focus of this project for several reasons. When geotagging, many users tweet simple location check-ins displaying statements such as "I'm at Palomino's in Los Angeles." In cases where one or both toponyms are Spanish words, the algorithm detects Spanish as the tweet language. The result is a massive over-representation of "Spanish" tweets in locations with Spanish place names even if other words in the tweet are written in English. Other language detection algorithms have been proposed and used with Twitter data, but none is perfect; Twitter's 140-character maximum combined with difficult to discern text (e.g., hashtags, informal language, and names) are limiting factors (Graham, Hale, & Gaffney, 2014). Second, assigning a language to a user based on individual tweets is difficult. Users can post in multiple languages and mix languages (some do frequently), but the user's account language likely reflects the language used for viewing other web content.

I focus on all users with an account language other than English rather than targeting a specific language, such as Spanish, for several reasons. Utilizing all non-English languages results in a greater sample size than any one language alone. After examining general patterns, more in-depth investigations on individual languages can be (and are) pursued. Additionally, the conventional race and ethnicity variables correspond to non-English users as a whole (e.g., percent white and percent foreign-born) rather than any single group.

Independent variables come from two U.S. Census sources: the 2014 American Community Survey (ACS) 5-year estimate and the 2014 Longitudinal Employer Household Dynamics Employment Statistics (LODES; see Table 1). ACS variables are residential and include population (POP), population density (POPDENSITY), median income (MEDINC), percent white (PERWHITE), median age (MEDAGE), and percent foreign born (PERFORBORN). Conversely, the working population is captured by the number of employees (JOBS) which comes from LODES. JOBS reflects a different land use pattern than POP, accounting for locations where people work but also buy goods and services. Many independent variables require calculations. POPDENSITY is calculated by dividing total population by land area. PERFORBORN is calculated by dividing the number of foreign born by total population. Since JOBS is only available at the block group level, values are aggregated to census tracts. Due to missing independent variable values, one census tract was omitted, reducing the dataset to 785 tracts. The DV under study, NETU, is not normalized by population, since this would presuppose a strong relationship between it and Twitter activity. Due to the emphasis on assessing the *data* in this project (as opposed to the focus on prediction), using a raw dependent variable is the preferred approach. These issues, among others, are discussed at length in later sections.

After collecting data, tweets from four automated accounts were removed: infosrv, which frequently posts server statuses; kartenquizde, which posts geographic quizzes as geotagged tweets; supra1Bqteam, which posts job advertisements; and trendinaliaHOU, which

posts currently trending topics in the Houston area. The geolocated tweets are stored in an open source NoSQL system, Elasticsearch (2017), and the counts of NETU in each tract are computed using Elasticsearch's geofilter query. A cardinality aggregation, which ensures each user is only counted once, is applied to the "user.id" field on each tweet. This prevents particular users who tweet often from inflating counts of the DV within single tracts. Queries are carried out with the Elasticsearch client for Python, and subsequent spatial analysis and statistics are completed using the R Project for Statistical Computing (R Core Team, 2017). Specifically, the "GISTools" package (Brunsdon and Chen 2014) is used for mapping, the "spdep" package (Bivand and Piras 2015) is used for SAR and tests of spatial dependence, and the "spgwr" package (Bivand and Yu 2017) is used for GWR. With each regression model, an ad hoc assessment of regression assumptions is carried out. Residuals are tested for normality and homoscedasticity using graphical methods, and a test of spatial autocorrelation (SAC), Moran's *I*, is used to test for spatial dependence in residuals. For Moran's *I*, a variable number of nearest neighbors are used in an attempt to identify spatial dependence at multiple scales.

Prior to the executing regression models, it was expected that POP, POPDENSITY, PERFORBORN, and JOBS would positively affect NETU while MEDAGE, PERWHITE, and MEDINC would negatively affect NETU. Hecht and Stephens (2014) have found urban biases of content production, indicating that areas with larger populations may have greater rates of LBSM usage. Since U.S. ethnic enclaves historically have had high population densities, higher rates of NETU usage in high population density zones are expected. The vast majority of Houston's immigrants speak a language other than English (Capps et al., 2015). From this it would follow that they would also be involved in networks that use a language other than English and use the web in an alternative language as well. Additionally, because geotagged tweets are commonly used for location check-ins at retail locations (i.e. locations with many employees), it was expected that JOBS would positively affect NETU.

Since younger populations have higher rates of social media usage (Zickuhr, 2013; Greenwood et al., 2016), MEDAGE is expected to have a negative relationship with NETU. Though smartphones and laptops are owned by a large percentage of the U.S. population, income nonetheless influences the likelihood of owning and being skilled with electronics (Rainie & Perrin, 2017). Therefore, the expected effect of MEDINC on NETU was positive. Finally, many of those speaking a language other than English in the U.S. are racial minorities; therefore, PERWHITE was expected to negatively influence NETU.

Preliminary analyses

Prior to executing various regression models, it is useful to examine some aspects of the raw data, such as the number of tweets/tracts per user, popular languages, and distribution of independent variables. In total, 26,354 tweets were produced by 5,693 NETU in the study area over the course of the data collection period. With the proposed scheme, users could potentially be counted in more than one tract, so it is also useful to examine the number of tracts each user tweeted from. The vast majority of users (68.7%) have tweeted from only one tract; 13.4% of users have tweeted in two tracts, 6.2% in three tracts, 3.3% in four tracts, and 2.1% in five tracts. Those who have tweeted in ten or more tracts comprise 2.0% of the dataset, and those who have tweeted in 20 or more tracts comprise 0.6% of the dataset. The average number of tracts per user is 2.1 with a standard deviation of 3.6. Users with Spanish as their account language dominate the study area, accounting for 72.6% of NETU, followed by Portuguese (6.6%), Japanese (3.9%), Turkish (3.8%), and Indonesian (2.5%; see Table 2).

While variable normality is not a regression assumption, it does provide a hint on a model's potential performance. Many variables are highly skewed, including the dependent variable, NETU (Fig. 2). Unsurprisingly, the initial OLS model reveals a failure to adhere to regression assumptions, thus requiring two modifications. Eight tracts have far and away the greatest NETU, to the point that they severely and adversely affect regression models. These tracts have NETU values of 1505, 1490, 844, 471, 355, 304, 239, and 112. The next highest

value is 79, followed by two values of 78. The eight greatest NETU tracts are well outside the typical upper bound for outliers, Q3 + 1.5*IQR, (with Q3 = 10, and IQR = 8). However, excluding all outliers by this criterion results in an omission of nearly 10% of the dataset. Since the gap between 79 and 112 is a reasonable natural break, the eight greatest tracts were deemed outliers and separated from the rest of the dataset. Rather than exclude these outliers completely, they are analyzed separately and reviewed at length in the discussion section.

Following the removal of outliers, OLS still performs poorly. Therefore, a transformation was sought for the DV. After evaluating candidate transformations, a Box-Cox (Box & Cox, 1964) transformation was applied. Using an iterative procedure, a series of values of λ are tested using Equation 1:

$$y^{(\lambda)} = \frac{y^{\lambda}-1}{\lambda} \quad \text{(Eq.1)}$$

For each value of λ, an OLS regression model is carried out with the new value of *y* and is assessed for performance using log-likelihood. Since zero values of NETU result in division by zero, each value of NETU is added to one, making all values positive. The λ value resulting in the greatest log-likelihood is then selected (in this case λ = -0.06). This value becomes the exponent of the original DV (NETU), resulting in a new DV (NETUTRANS). Non-linear transformations can complicate interpretation, but in this case it is fairly straightforward. The largest values of NETU are the smallest values of NETUTRANS, and vice versa. Strong relationships with NETU typically result in a similarly strong relationship with NETUTRANS but with the opposite sign. However, this is not always the case, particularly when original relationships are weak. Every subsequent model described uses NETUTRANS as a DV as opposed to NETU.

Results

The OLS model with the transformed DV reveals that all independent variables are significant except POPDENSITY. These results must be considered with caution due to failures in meeting

regression assumptions. While this model improves upon OLS using the non-transformed DV, it does possess residual heteroscedasticity (see Fig. 3). Additionally, Moran's *I* shows significant SAC at every scale tested, from 4 to 92 nearest neighbors in increments of 8 (Table 3). These shortcomings call for SAR models, which more effectively compensate for the spatial structure of the data.

After testing several bandwidths, 20 nearest neighbors were selected for the spatial lag and error models. The SAR models result in significant improvements in terms of AIC and log-likelihood (see Table 4). However, they show only marginal improvements (if any) over OLS in terms of residual normality and heteroscedasticity Like OLS, the spatial error model shows highly significant SAC at every bandwidth tested, and the spatial lag model demonstrates significant SAC with greater than 20 neighbors.

GWR was subsequently pursued due to the shortcomings of these models and the desire to find locations of non-stationarity among independent variables. Using a golden section search to determine a bandwidth of roughly 10 km, a bi-square kernel was implemented. The GWR model demonstrates several notable findings. Every variable exhibits some positive and negative standardized coefficients, with the exception JOBS (Table 5 and Fig. 4). With this variable, all coefficients are negative, but each other variable has a positive effect on NETUTRANS in some locations and a negative effect in others. The presence of *significant* negative and positive standardized coefficients is found only with PERWHITE. This variable shows one large cluster of negatively significant values in the central and southwest portion of the county and one small cluster of positively significant values toward the Northeast. PERFORBORN and MEDINC demonstrate interesting patterns. In most locations, their effect is insignificant, but in one small cluster their effect is strongly and negatively significant. The spatial patterns of standardized coefficients is less interesting for the other variables. POP and JOBS exhibit almost exclusively negative effects, MEDAGE's effect is mostly positive, and POPDENSITY's effect is mostly insignificant. Particular aspects of nonstationary clusters are

explored in detail in the discussion section.

Discussion and post hoc analyses

GWR shows that JOBS clearly has the greatest effect on NETUTRANS, demonstrating that users prefer to geotag in areas with many employees, bustling with activity. PERWHITE, on the other hand, has a significant effect on NETUTRANS but mostly in the opposite direction expected. Its effect on NETUTRANS is negative (meaning it has a positive effect on the non-transformed version of the variable, NETU) in most of its significant areas. Since the presence of languages other than English is often associated with non-white populations, this effect is counterintuitive if users prefer to tweet from home. However, as shown by Haffner et al. (2017), the vast majority of users prefer to geotag posts away from home. The outliers further shed light on this facet, but discussion is first devoted to the non-stationary clusters identified in GWR.

The two significant PERWHITE clusters (see Fig. 5) are easily explainable when examining aspects of the raw data. The positively significant cluster, in northeast Harris County overlaps with Kingwood, Texas. The tracts in this cluster have higher proportions of white population (most are greater than 90%), have high median income, and possess few NETU. These tracts collectively contain only fifteen users: thirteen using Spanish, one using Portuguese, and one using Turkish (Table 6). The negative cluster, on the other hand, covers a much larger portion of the city. It has an over-representation of Spanish NETU (80.0% of NETU in these tracts as compared to 72.6% in the county as a whole), and contains a mix of high and low values for independent variables. The reason for a lack of NETU in the Northeast is twofold: due to its demographics and high income, it is likely dominated by English speakers, and the area is mostly residential, lacking notable check-in locations. The cluster of tracts where PERFORBORN negatively affects NETUTRANS (toward the South) contains a mix of both high and low raw values for MEDINC, PERFOBORN, and JOBS. This area is largely non-white. Interestingly these tracts have a disproportionately large number of Turkish NETU (10.5%) compared to the rest of the county (3.8%) though the reason for this is undetermined.

The outliers collectively tell a compelling story. These eight tracts (Fig. 6), possessing the greatest NETU in the county, each contain a notable amenity and are predominantly non-residential (Table 7). The tract with the greatest NETU contains George Bush International Airport, accounting for 13.0% of all NETU in the county. The vast majority of tweets in this tract are location check-ins with text such as "I'm at Gate B14 in Houston, TX." The tract with the second greatest NETU is just west of Downtown and contains Eleanor Tinsley Park, a green space area used for outdoor concerts. The third greatest NETU tract is Downtown Houston, containing many attractive locations such as Minute Maid Park (home of the Houston Astros), the Toyota Center (a concert venue and general event space), and a wealth of restaurants, hotels, and civic buildings. The fourth greatest NETU tract lies in the far southeast portion of the county, home to Space Center Houston. The fifth greatest tract houses the Galleria, a large mixed-use shopping area. Other tracts in the top eight have similar notable features: NRG Stadium (home to the Houston Texans), NRG Arena (a multi-purpose event facility), the Houston Zoo, and other shopping centers.

Independent variables vary wildly within the top eight NETU tracts. The tract with the greatest number of NETU, home to George Bush International Airport, has the fourth lowest POP in the dataset. Though not nearly as low as this tract, most other outliers are in the bottom half of the dataset by POP. This is notable given that POP negatively affects NETUTRANS in the spatial lag model, thus positively influencing NETU. The tract with the greatest NETU also has a very low MEDINC, in the bottom 25%, while all other top NETU tracts are in the top half by MEDINC with most in the top 25%. Similarly, tracts ranking second and third by NETU vary drastically in terms of JOBS: the second greatest NETU tract ranks near the bottom 25% while the third greatest NETU tract has the greatest value for JOBS in the entire dataset: 154,338, accounting for 6.7% of all JOBS in Harris County. Such variability wreaks havoc on regression models.

Significant variability among independent variables in locations with the most Twitter

activity warrants a word of caution. For one, the inclusion of such tracts makes prediction exceptionally difficult. Beyond this, infrequent events – airline flights, concerts, sporting events, and shopping sprees – as opposed to day-to-day interactions, appear the primary drivers of content production for users in this dataset. Additionally, considering that 68.7% of users only posted in one census tract and that the tract with the greatest NETU houses the International Airport, it is reasonable to suspect that many of these users are not residents of Harris County. For these reasons, residential characteristics of places should not be ascribed to users producing content in these places. This finding suggests that non-English geolocated Twitter data says much less about users' urban dynamics and more about the digital status of locations. In this way, place is conceptualized as a "hybrid reality" (de Souza e Silva, 2006) – a coalescence of the physical aspects that make a place attractive for geotagging combined with its resulting digital visibility. That said, non-stationarity among several independent variables identified through GWR should caution researchers against making sweeping statements about the types of locations users prefer. The effect of these variables could vary greatly both within and between other cities.

Originally, it was determined that normalization would be necessary since NETU would likely be skewed in favor of high population tracts (i.e. a greater presence of people leads to more Twitter activity). While POP is highly significant in most tracts in GWR, the relatively low POP among the high NETU outliers shows that normalization by residential population would be a poor idea. If any normalization is to be applied, it would be better to normalize by JOBS or a similar variable, such as number of retail establishments. Informal observations of tweet text show that a large proportion of this dataset is comprised of location check-ins where users visit such locations. This importantly calls into question the nature of spatial information in precisely located tweets. What was perhaps considered ambient geospatial information may not be so ambient after all; the location check-in is more explicit, integrating spatial information with a post's text. The use of general location on Twitter, not to mention other forms of LBSM, may be different.

In some ways it seems counterproductive to utilize precise location Twitter data only to aggregate to a coarser spatial resolution. However, joining general location data to a specific geography is much more difficult. General location data returns a place name and rectangular bounding box defined by four points. These can vary greatly in size within and between location types. A neighborhood's bounding box, while generally smaller than other location types, may or may not be contained by a single census tract. Conversely, city bounding boxes can intersect multiple counties and potentially many other cities. A user can be present at any location within the bounding box – or completely absent from the bounding box if they fabricate their location – making it impractical to use traditional census geographies with general location. At the extreme, a user can tag a country, which is of little use in most circumstances. Yet, language processing methods, like those put forth by Longley et al. (2015) and Stefanidis et al. (2013), could be used on tweets with general location (or no location) to estimate users' locations.

The effects of an imposed scale on this dataset cannot be neglected. At some scales, the prevalence of Twitter activity seems to closely follow population patterns. Global maps of all Twitter activity presented by Leetaru et al. (2013) reflect the world population distribution, albeit with an over-representation in Europe, the United States, and Japan. Similarly, Twitter activity symbolized by language in Western Europe closely aligns with international borders. At the county level in Texas, the counties with the greatest NETU generally have the greatest population, but this pattern is not followed at the census tract level as described earlier in the discussion section. At coarse scales LBSM may very well be reflective of offline processes, but at finer spatial scales perhaps not. Other aggregations that are finer still (e.g., block groups and block) would be much sparser but could potentially reveal other, yet unknown patterns.

The approach taken in these analyses is not without drawbacks. While someone could create an unlimited number of posts in one census tract and still only be counted as one user, limits were not imposed on the number of census tracts a user could post in. Given that 2.0% of users posted in ten or more tracts, it is apparent that this method does not completely

eliminate the long-tail effect. Additionally, this study does not capture all Twitter users who speak (or possibly prefer) a language other than English. For various reasons a user may choose to receive emails and notifications from Twitter in English while consistently posting in a different language. Other criteria could be applied to capture these users more effectively, such as requiring that users post in another language 50% of the time or more. This analysis also neglected temporal effects. While a pilot study using less data (roughly ten months) did not reveal any significantly different findings, dividing the data into meaningful periods (e.g., based on various holidays, sports seasons, or election schedules) could reveal unique spatial patterns.

Despite this study's drawbacks and words of caution, there are several positive findings. It has revealed popular locations within Houston and established that precisely located Twitter data says more about places than users. Additionally, the top eight NETU tracts differ somewhat from the top eight total Twitter users' (which includes those using English) tracts. While English account users were not the focus of this study, this difference could be indicative of meaningful, yet subtle, spatial differences between English and non-English users. This study also demonstrates an effective use of personal profile account information in eliciting spatial differences as opposed focusing on tweet text. A wealth of other account information is available, such as a user's number of followers, number following, profile description, and more.

Conclusion

This study demonstrates that POP, MEDAGE, and JOBS are the most influential on NETU in Harris County. The non-residential variable, JOBS, has a stronger effect than any residential variable. However, each regression model conforms to assumptions poorly, and GWR reveals significant patterns of non-stationarity. Results would likely be different in other counties and possibly during other time periods. Additionally, residential variables are not representative of Twitter users and should be understood as merely characteristics of those locations. In line with this, this subset of LBSM at this scale says more about digital status of locations than the day-

to-day patterns of users. It reveals popular locations for geotagging, but this does not necessarily mean that users frequent these locations most often. Some of the most popular locations for geotagging – George Bush International Airport, Space Center Houston, and the Houston Zoo – are locations that users likely only visit intermittently, if not rarely. For these reasons, precisely geotagged tweets are likely not representative of users' home locations nor day-to-day travel patterns. Observations of users who tweet from multiple locations within a small time period (e.g. four hours), as utilized by Lee et al. (2016), may be more representative of users' everyday lives. Yet, this activity is generated by such a small cohort of users that the capability for generalization with these methods should be questioned.

These findings demonstrate the difficulties in exposing important patterns of language use and prohibit drawing any definitive conclusions about race and ethnicity. While these findings should invoke caution against generalization and reinforce the importance of heterogeneity and spatial context (Miller, 2017), this study is not a complete diatribe against using LBSM to study spatial processes. The apparent gap between residential characteristics and LBSM perhaps offers opportunities for building new geographic knowledge. Other important findings are noted but not yet fully understood, such as the relative abundance of Turkish users and the differences between NETU and Twitter users more generally. This study also demonstrates an effective use of account information as opposed to focusing on LBSM text, and other studies could take advantage of variations on this approach. Additional information on user behavior could lead to appropriate applications, but other work must continually build upon our current understanding of LBSM data.

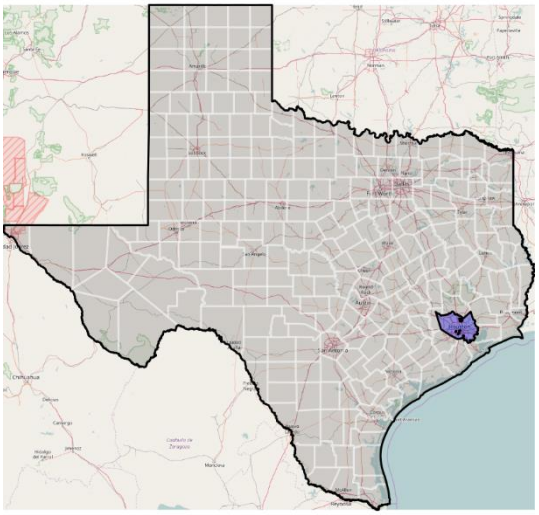

Figure 3.1 Location of Harris County within Texas

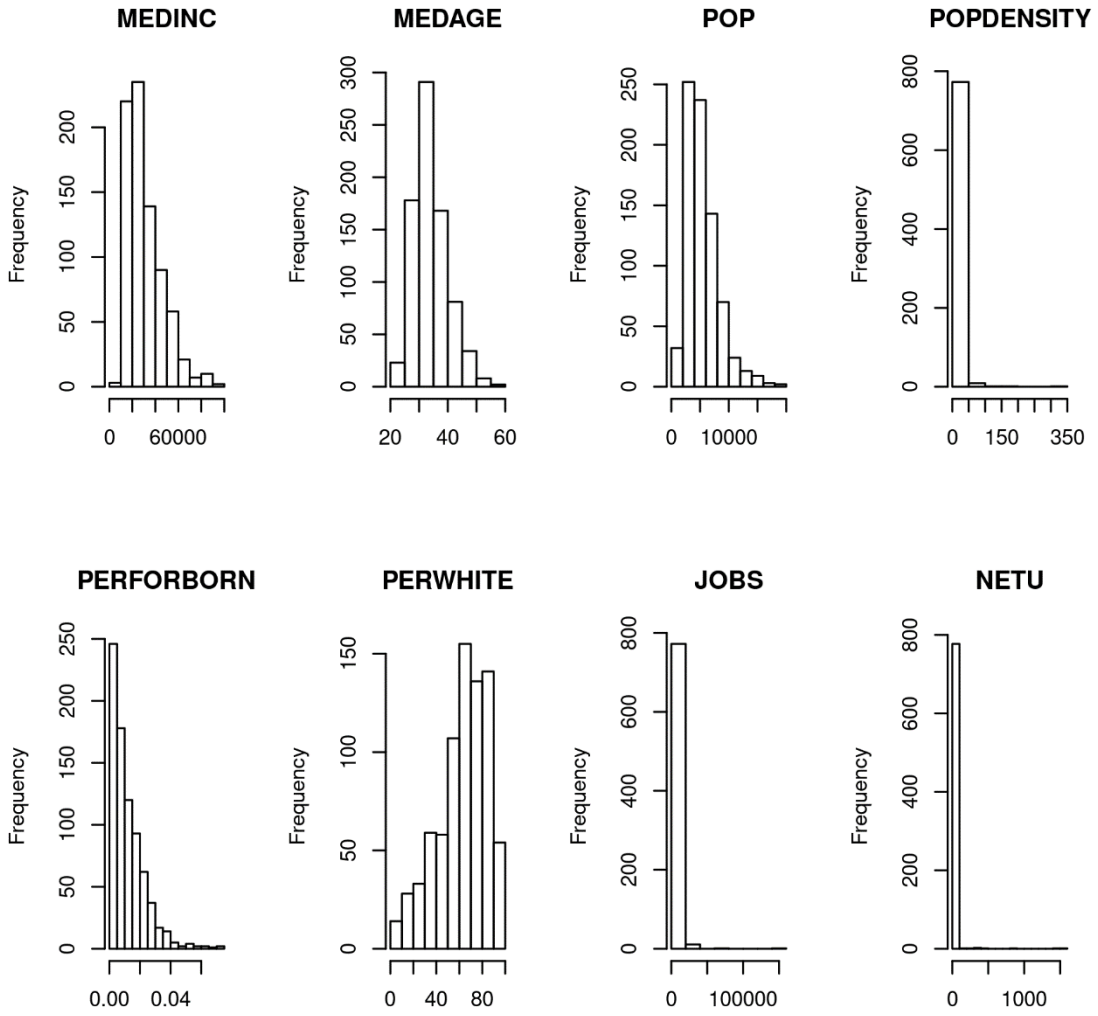

Figure 3.2 Histograms of raw variables

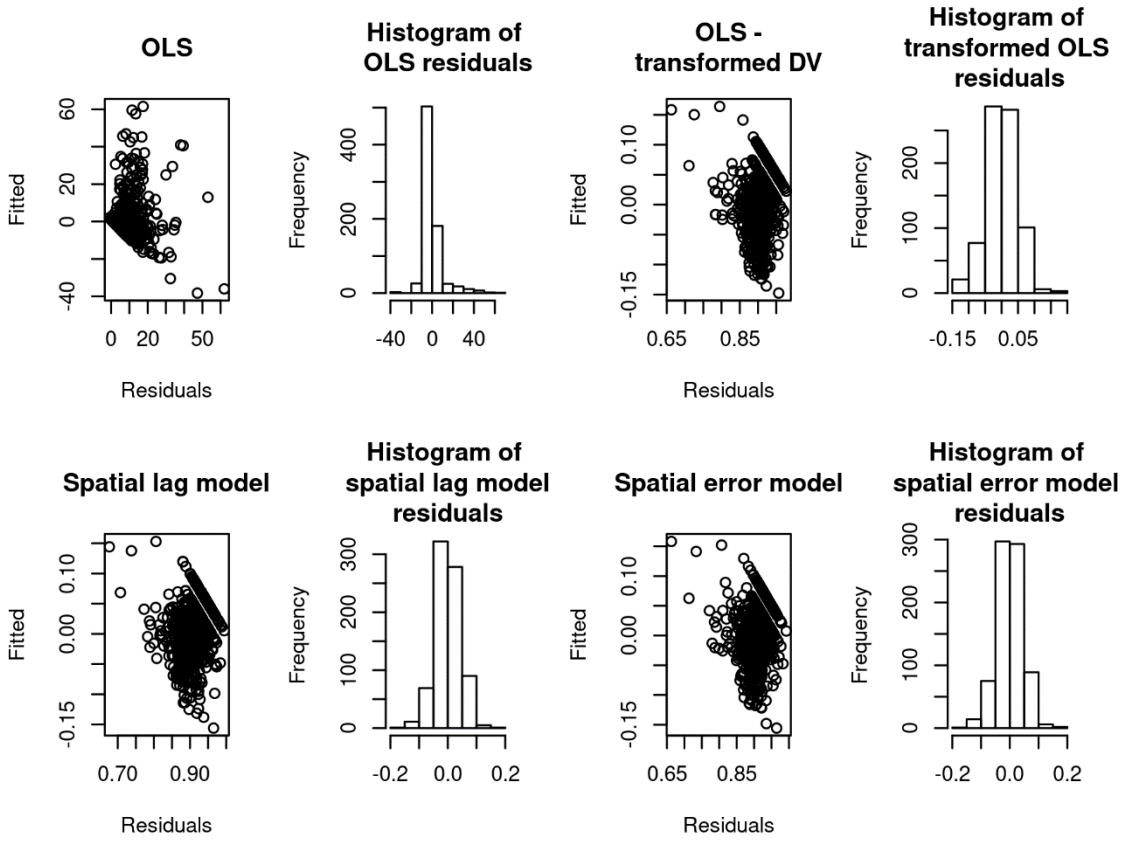

Figure 3.3 Regression diagnostics: histograms of residuals and fitted vs. residual values

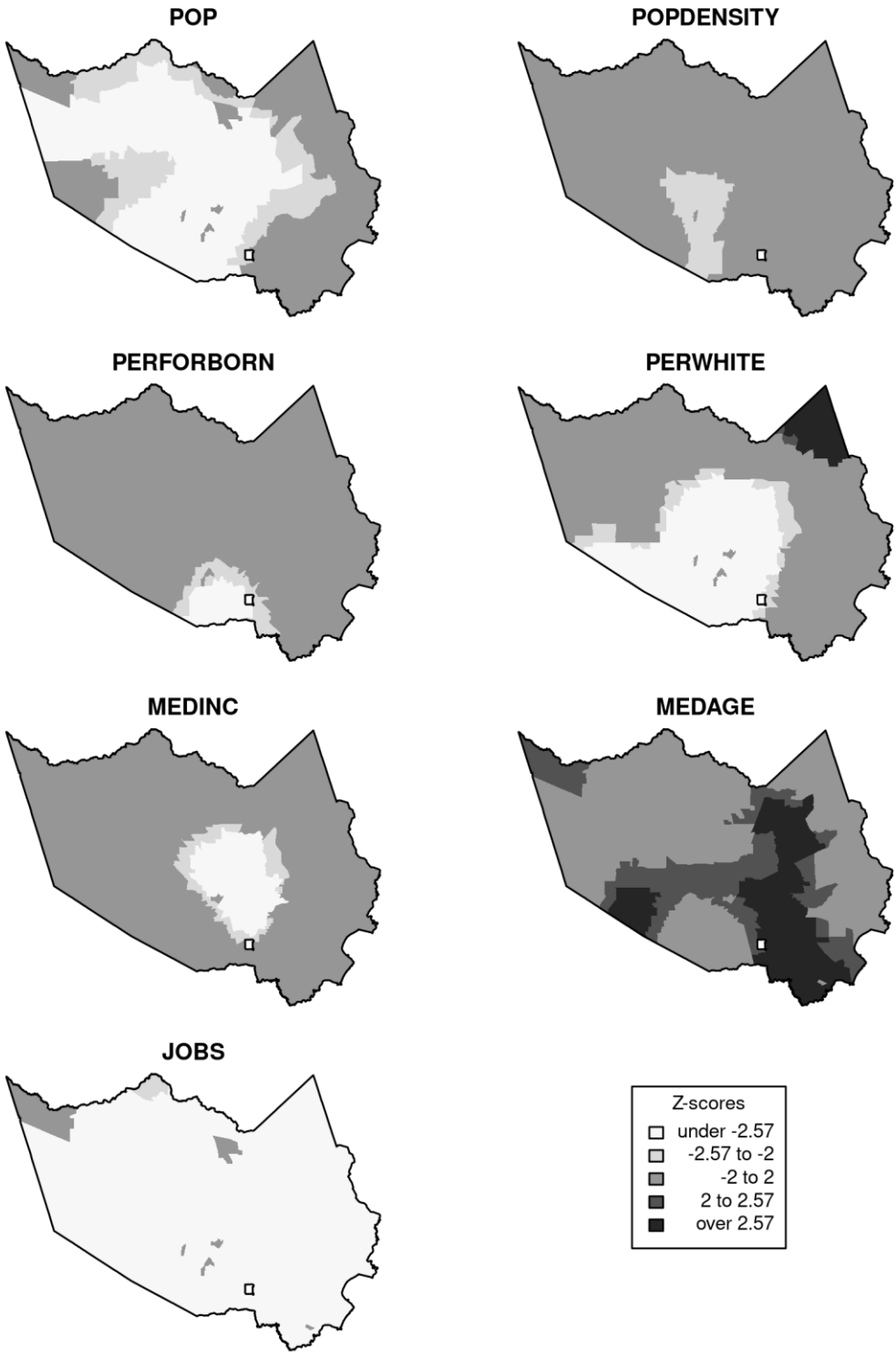

Figure 3.4 Standardized GWR coefficients

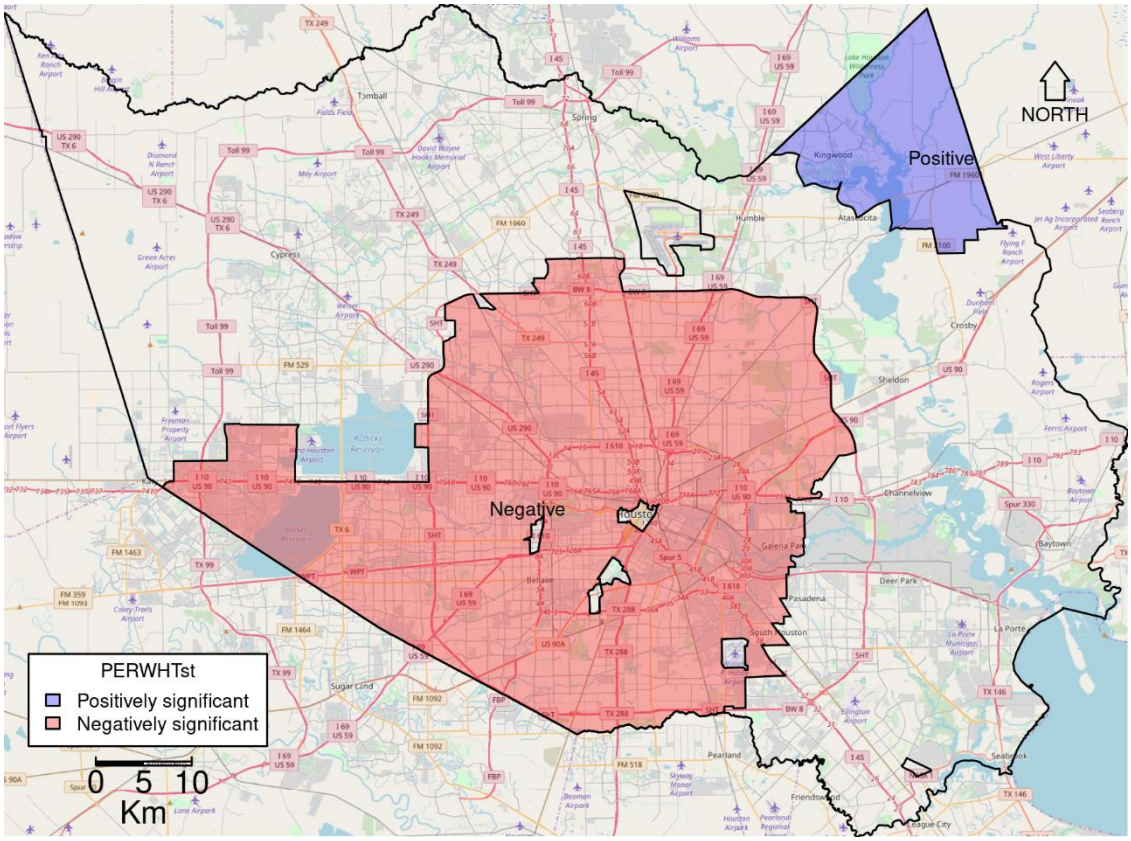

Figure 3.5 Clusters where PERWHITE has a significant effect on NETUTRANS

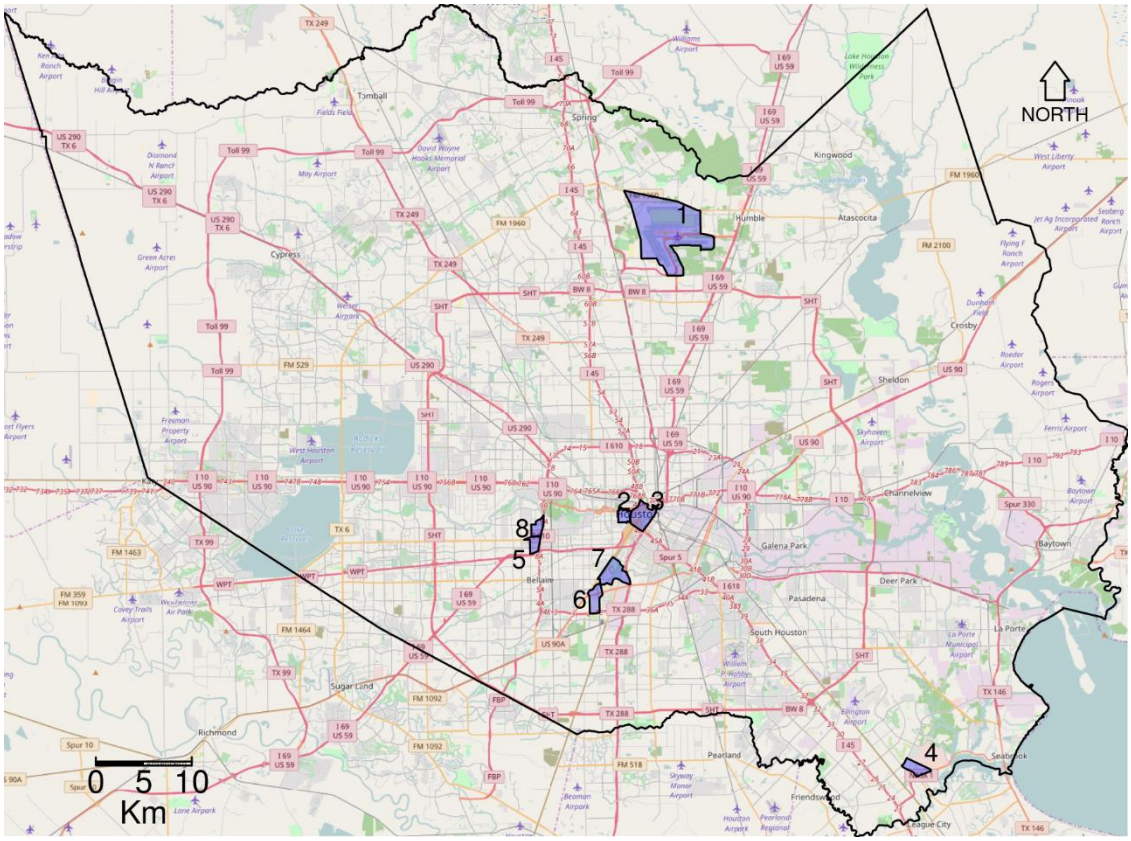

Figure 3.6 Locations of NETU outliers

Table 3.1 Regression Variables

Variable name	Code	Type	Source	Calculation	Expected relationship with DV
Population	POP	Independent	2014 American Community Survey (5-year estimate)	None	+
Population density	POPDENSITY	Independent	2014 American Community Survey (5-year estimate)	Population divided by land area	+
Median income	MEDINC	Independent	2014 American Community Survey (5-year estimate)	None	+
Median age	MEDAGE	Independent	2014 American Community Survey (5-year estimate)	None	-
Percent foreign born	PERFORBORN	Independent	2014 American Community Survey (5-year estimate)	Number of foreign born divided by population	+
Percent white	PERWHITE	Independent	2014 American Community Survey (5-year estimate)	None	-
Number of employees	JOBS	Independent	2014 Longitudinal Origin-Destination Employment Statistics	Spatial join of blocks inside census tracts	+
Number of non-English Twitter users	NETU	Dependent	Streaming Twitter API, Oct. 2015 - Nov. 2016	Number of non-English Twitter Users	N/A
Number of non-English Twitter users transformed	NETUTRANS	Dependent	Streaming Twitter API, Oct. 2015 - Nov. 2016	NETU^λ	N/A

Table 3.2 User language counts in Harris County

Rank	Language	Number of users	%
1	Spanish	4133	0.726
2	Portuguese	377	0.066
3	Japanese	223	0.039
4	Turkish	216	0.038
5	Indonesian	142	0.025
6	French	139	0.024
7	Italian	90	0.016
8	Arabic	84	0.015
9	Russian	60	0.011
10	German	54	0.009

Table 3.3 Moran's *I* on residuals

Neighbors	OLS (transformed)		Spatial lag model		Spatial error model	
	statistic	p-value	statistic	p-value	statistic	p-value
4	0.203	0.000	0.034	0.062	0.131	0.000
12	0.176	0.000	0.012	0.164	0.102	0.000
20	0.152	0.000	0.003	0.342	0.080	0.000
28	0.139	0.000	0.013	0.041	0.071	0.000
36	0.122	0.000	0.016	0.013	0.057	0.000
44	0.111	0.000	0.016	0.006	0.047	0.000
52	0.097	0.000	0.012	0.017	0.035	0.000
60	0.090	0.000	0.013	0.006	0.031	0.000
68	0.084	0.000	0.014	0.002	0.027	0.000
76	0.076	0.000	0.012	0.005	0.021	0.000
84	0.070	0.000	0.011	0.005	0.016	0.000
92	0.065	0.000	0.010	0.006	0.011	0.002

Table 3.4 Model diagnostics

Model	AIC	Log likelihood	Spatial term significance
OLS (transformed DV)	-2536	1277	NA
Spatial Lag	-2642	1331	0.000
Spatial Error	-2582	1301	0.000

Table 3.5 Standardized GWR variable ranges and R-squared

Variable	Minimum	Median	Max	Std. Dev.
POPst	-4.156	-3.097	0.198	1.0359
POPDENst	-2.396	-0.958	0.481	0.7496
PERFOst	-3.586	-0.854	0.826	0.8247
PERWHTst	-6.013	-3.144	3.475	2.1497
MEDINCst	-4.503	-0.900	1.885	1.3791
MEDAGEst	-0.436	1.956	3.731	0.9807
JOBSst	-9.369	-7.609	-1.926	1.5476
localR2	0.246	0.392	0.624	0.0784

Table 3.6 Notable languages within select clusters of significant GWR coefficients

Significant cluster	Spanish users	%	Portuguese users	%	Japanese users	%	Turkish users	%	Total NETU
MEDINCst cluster (-)	1073	0.838	37	0.029	25	0.020	50	0.039	1281
PERFORBORNst cluster (-)	579	0.758	26	0.034	12	0.017	80	0.105	764
PERWHITEst cluster (-)	3401	0.800	139	0.033	90	0.021	310	0.073	4273
PERWHITEst cluster (+)	13	0.867	1	0.067	0	0.000	1	0.067	15

Table 3.7 NETU Outliers and independent variables

NETURank	NETU	POP	POP DENSITY	PERFOR BORN	PER WHITE	MEDINC	MEDAGE	JOBS	Significant amenity
1	1505	787	0.276	0.000	35.600	17574	26.8	23973	George Bush International Airport
2	1490	3828	11.427	0.018	45.300	42419	30.5	573	Eleanor Tinsley Park
3	844	4178	1.249	0.011	73.000	51063	43.6	154338	Downtown Houston (Minute Maid Park, Toyota Center, other amenities)
4	471	2984	1.979	0.027	67.200	31445	30.5	5839	Space Center Houston
5	355	3250	7.303	0.031	72.800	57350	36.7	26701	The Galleria
6	304	3451	1.744	0.008	43.200	41352	29.1	14516	NRG Arena and NRG Stadium
7	239	2976	0.875	0.011	59.000	44117	31.2	72932	Houston Zoo
8	112	3472	5.701	0.026	71.900	49066	34.1	27608	The Galleria

CHAPTER IV

A PLACE-BASED ANALYSIS OF #BLACKLIVESMATTER AND COLOR-BLIND RACISM ON TWITTER[3]

Introduction

Blacks and whites in the United States have long had polarized perceptions on race-related political issues (Massey and Denton 1998). Recent reactions to the string of shootings of unarmed black men in tandem with the 2016 presidential campaign season, which culminated with the election of Donald Trump, have exemplified this already stark divide. Social media, particularly Twitter, has become a public ground for debate on racial issues where hashtags such as #BlackLivesMatter and others are used for protest and counter-protest. Research on such topics is valuable because it sheds light on the formation of collective identities (Ray et al. 2017), how the web is used to demonstrate (and build) solidarity (Ince et al. 2017), and how racism can be combated (Byrd et al. 2017). While nascent literature has documented social characteristics of this online debate, research on its spatial manifestation is lacking. In this article, I undertake a place-based approach to studying the factors driving #BlackLivesMatter and counter-protest content (e.g., #AllLivesMatter) in Louisiana and Texas cities, discussing how patterns correspond to conventional data sources, how results fit into germane racial theory, and implications of contributed geographic information using text and profile content.

[3] Currently under review as: Haffner, M. (2018). A place-based analysis of #BlackLivesMatter and Color-blind racism on Twitter.

Hashtag activism, #BlackLivesMatter, and color-blind racism

The use of hashtags on social media has largely been successful in projecting the voices of oppressed groups and starting political movements (Fekete 2018). By keeping running totals of commonly used hashtags and providing links to posts containing that content, social media platforms allow users to quickly engage with popular content. The widespread nature of the #OccupyWallStreet hashtag resulted in offline protests in September 2011, which in turn knit together a community of individuals prepared to participate in disaster relief efforts of Hurricane Sandy through #OccupySandy (Donovan 2015). Though not directly causing political uprisings in the "Arab Spring", hashtag activism did indeed significantly contribute to the destabilization of Middle East governments (Fekete and Warf 2013).

Blacks have been particularly active on Twitter as evidenced by high penetration rates of this demographic group in the U.S. (Pew Research Center 2018) and the cohesive, yet complex, community surrounding "Black Twitter" (Clark 2014). The acquittal of George Zimmerman, who fatally shot black 17-year-old Trayvon Martin in 2012, sparked the birth of the #BlackLivesMatter movement, raising awareness about racial injustice and police brutality (BlackLivesMatter 2018). First appearing on Facebook, the phrase has rapidly proliferated across American culture (Ince et al. 2017). Today, #BlackLivesMatter continues to be especially prominent on Twitter and has intensified after the deaths of unarmed blacks at the hands of police – notably Michael Brown, Freddie Gray, John Crawford, Ezell Ford, Philando Castile, and Alton Sterling. The persistence of #BlackLivesMatter over time is unprecedented compared to other hashtags used for social causes. It was the seventh most used hashtag globally on Twitter in 2016 (Berland 2016) and seventh most used in the U.S. in 2017 (Machin 2017), despite first appearing years earlier. While other hashtags, such as #AltonSterling, #SayHerName, and #Ferguson, emerged after specific events, #BlackLivesMatter encapsulates a movement and functions as a canonical hashtag for a plethora of race-related developments.

As a method of counter-protest to #BlackLivesMatter, #AllLivesMatter emerged on

Twitter, albeit in a more disorganized fashion. The appearance of this latter hashtag is a prime exhibition of what Bonilla-Silva (2006) calls "color-blind racism". Bonilla-Silva (2006) contends that while most whites say they agree with civil rights principles, they do not support them in practice. They claim to not see race, which paradoxically reinforces subversive racist policies and practices. In an analysis of the 1998 Detroit Area Study and the 1997 Survey of Social Attitudes of College Students, he uncovers four themes used to downplay the effects of race in inequality: (1) abstract liberalism, in which respondents use vague references to individualism, hard work, and meritocracy; (2) naturalization, in which respondents claim that "that's just the way it is" (p. 37); (3) cultural racism, in which respondents suggest that blacks are lazy or culturally deficient, thereby "blaming the victim" (p. 40); and (4) minimization of racism, in which other factors, such as class, are made to be more important than race. Some use the minimization theme to ignore the effects of discrimination outright, which is similarly reflected in the general message of #AllLivesMatter. While strands of each frame – and combinations of frames – can be found accompanying references to #AllLivesMatter, it fits the minimization theme the best. The proponents of #AllLivesMatter actively discount the disproportionate amount of police brutality against minorities as a systemic issue (Rios 2011; Alexander 2010).

Contrary to common perceptions, Ray et al. (2017) find that #TCOT (Top Conservatives on Twitter) was the primary counter-narrative to #BlackLivesMatter in the wake of Michael Brown's death in August 2014. In fact, during this time #TCOT was used much more often than #AllLivesMatter and contained many counter-protest themes including 'validating justifiable homicides,' 'humanizing police officers,' and 'white victims of black criminality.' Similarly in January 2015, #TCOT was used with similar frequency to that of #BlackTwitter (Graham and Smith 2016). This seemingly non-racial #TCOT hashtag – beyond the more oppositional #AllLivesMatter – speaks to the subversive nature of color-blind and systemic racism.

On average, perceptions of blacks and whites sharply diverge on issues of race, equality,

and desires for diversity. Forty-one percent of blacks in the U.S. strongly support #BlackLivesMatter versus only fourteen percent of non-Hispanic whites (Horowitz and Livingston 2016). Thirty-eight percent of non-Hispanic whites feel that "our country has made the changes needed to give blacks equal rights with whites" versus only eight percent of blacks (Horowitz and Livingston 2016, 4). As pointed out by Warren (2011), the literature on perceptions of police by race is extensive (see Brunson 2007; Weitzer and Tuch 2006; 2005; 2004; Brandl et al. 1994), with blacks predominantly viewing police more unfavorably. According to the 2004-2005 Chicago Area Study, whites prefer less diverse neighborhoods than blacks and Latinos, and when searching for a house they explore neighborhoods much less diverse than their preferences indicate (Havekes et al. 2016).

The skewed preferences – combined with a multitude of structural factors, such as racial steering in home buying, restricted access to home loans through redlining, and perpetual concentrations of poverty preventing upward mobility – have led to stark contrasts in racial residential alignment (Massey and Denton 1998). Despite legislation attempts to reduce segregation in the mid to late 20th century, residential racial patterns have merely been reconfigured (Ellis et al. 2017). The outcome has ensured social exclusion leading to racial inequalities in education, income, wealth, access to health care, and virtually all other areas of life.

Space, place and race on the GeoWeb

Recent studies on volunteered geographic information (VGI; Goodchild 2007) content illuminate spatial differences related to present-day segregation. Like the decentralized nature of #BlackLivesMatter as a social movement (Ince et al. 2017), spatial information is similarly produced through VGI. Through a variety of platforms and subtypes (e.g., open mapping projects, citizen science initiatives, location-based social media or LBSM) users can contribute locational information at will, participating in the social construction of the GeoWeb. On the surface, these projects would appear to function as a form of liberation for those historically excluded from the technological realm, yet gendered (Stephens 2013; Haffner et al. 2017) and

racialized (Crutcher and Zook 2009; Fekete 2015) power structures work to keep the GeoWeb uneven.

In the aftermath of Hurricane Katrina, the Lower Ninth Ward, a predominantly black residential area, was substantially lacking in user-created Google Earth placemarks compared to the rest of the city (Crutcher and Zook 2009). Similarly, Fekete (2015) finds a vast underrepresentation of black neighborhoods on the check-in platform, Foursquare, across multiple U.S. cities. In these cases, multiple factors are responsible for spatial imbalances. Both a 'digital divide' (Warf 2001) – that is, a lack in physical access to devices used to participate in the GeoWeb – and a lack of desirable locations in physical space (Fekete 2015) produce uneven patterns on the web. While a lack of digital content is not disparity per se, it is reflective of the offline processes driving inequality. Finding more mixed results, Shelton et al. (2015) detail patterns of geotagged Twitter usage near the '9th Street Divide' in Louisville, Kentucky. This imaginary boundary between the more affluent, predominantly white East End and the lower income, predominantly black West End is commonly believed to be strict. However, words such as 'ghetto' appear in geolocated tweets in the West End, around the West End, and throughout Louisville, thus complicating discourse on 9th Street Divide. Additionally, users who predominantly tweet from the West End appear more mobile in that they tweet from a greater variety of locations throughout the city, despite the fact that residents in this area have lower incomes on average (Shelton et al. 2015).

Theoretical framework and research questions

As demonstrated by Shelton et al. (2015), the study of space and place through VGI has the potential to reveal more nuanced racial patterns than with conventional data sources alone (Graham and Zook 2011). While critiques of using big data in a geographic context have been numerous and are not without merit (e.g., Haffner 2018, Longley et al. 2015), the efficacy of these data sources in their ability to elicit patterns of inequality is often overlooked (Shelton et al. 2014). Additionally, big data need not be used in a theory-free, positivistic sense (Miller and

Goodchild 2015). A promising, and arguably more sound, approach is that of "abductive reasoning" – a type of inductive reasoning in which the goal is not the proof or disproof of hypotheses but the formation of hypotheses through data analysis (Miller 2010).

In this article, I utilize abductive reasoning (Miller 2010) in the context of color-blind racism (Bonilla-Silva 2006) to explore the relationship between racial protest through #BlackLivesMatter, counter-protest through other phrases, and residential demographic variables from the U.S. Census. Given the historic polarization between blacks and whites on racial issues, current residential segregation patterns, and the building body of literature suggesting offline processes are reflected online, it would be expected that geotagged #BlackLivesMatter tweets and corresponding counter-protest content would be similarly discursive: areas with higher percentages of blacks should exhibit more #BlackLivesMatter content, and areas with higher percentages of whites should reveal more counter-protest content. Using cities within Louisiana and Texas, I explore the spatial patterns of #BlackLivesMatter and counter-protest content, through cartographic visualization and ordinary least squares (OLS) regression. Specifically, I ask: (1) Which census variables best predict the production of #BlackLivesMatter and counter-protest content in Texas' and Louisiana's cities? (2) Which cities are outliers, and what do tweets from individual users tell us about protest and counter-protest? (3) More generally, how can data from Twitter inform us about socio-spatial processes?

Data and methods

All geotagged tweets in the U.S. were collected using Python and the Twitter streaming application programming interface (API) from 17 October 2015 to 26 November 2016. This API gives much more than a tweet's text; it returns a JavaScript object notation (JSON) array containing a user's screen name, profile description, self-defined profile location, number of followers and followees, hashtags used, and language used (both as defined by the user and estimated through Twitter's language detection algorithm) among other variables. If a user enables location on a tweet, the API also returns information such as the country where the

tweet was produced, the place name (usually a city, town, or point of interest) as selected by the user, and latitude-longitude coordinates of the location. Twitter users can either select 'precise location', which appears as an exact latitude-longitude pair as determined by the user's device, or general location, which allows the user to select from a list of FourSquare locations and is represented by a four-point bounding box around the location selected (see Fig. 1). These locations are suggested based on the user's current location, with nearer locations toward the top of the list. While difficult to fabricate a precise location, a user may select any general location, even one not listed in the default drop-down list. However, it is anticipated that most users do not intentionally geotag false locations, considering that individual online identities are typically consistent with offline identities; Kennedy (2006) finds that individuals are likely to divulge too much personal information on the internet rather than fabricate aspects of their identity, even when it would be expedient to do so. This study makes use of all geotagged tweets with "place_type" equal to "city", which includes both precise and general location tweets.

Municipalities within Louisiana and Texas were chosen as the study area for several reasons. The two largest metropolitan areas in both states – New Orleans, Baton Rouge, Houston, and Dallas – contain municipalities with large, disparate black and white populations. Additionally, significant, albeit different, race-related events occurred in both states during the data collection period. The death of Alton Sterling on 5 July 2016 in Baton Rouge, Louisiana, received national attention after a disturbing video of his death surfaced online. This caused a surge in #BlackLivesMatter tweets across the country but especially in Baton Rouge. Two days later, five police officers were killed in Dallas, Texas, causing a surge in counter-protest tweets. Not wanting to veil potentially interesting and unexpected patterns, all cities within these two states are first examined as opposed to isolating individual cities (such as only Dallas and New Orleans) a priori.

Tweets were originally collected and stored in flat files on hard disks. In preparation for analysis, tweets containing the desired phrases were parsed using a combination of the Bourne Again Shell (Bash) and Python. #BlackLivesMatter was determined to be the protest

text of primary focus, but both #AllLivesMatter and #TCOT were examined as potential counter-narratives. Objects containing the word(s) 'BlackLivesMatter', 'AllLivesMatter', and 'TCOT' (case insensitive and without the '#' character) anywhere within each JSON array were extracted. Subsequently, each tweet was indexed into its corresponding ElasticSearch index and queried using the Python Elasticsearch Client. Within indices, the data were further subdivided into only those tweets within Texas and Louisiana. Further analyses were completed with the R Project for Statistical Computing.

The data analysis focuses on two differing types of reference to these two phrases: text and profile. This is done, in part, in an attempt to differentiate between indications of attention versus support (see Graham and Zook 2011). For example, a user can make a negative text reference to #BlackLivesMatter in the text of a tweet, but it is theorized that the inclusion of the phrase #BlackLivesMatter in a user's profile more likely signifies support. The number of users referencing each phrase was counted within each place (i.e. municipality) in Texas and Louisiana rather than counting individual tweets with the goal of eliminating the long-tail effect – the adverse consequence of a small number of individuals producing a disproportionately large amount of content (Elwood et al. 2013). Counts within each place were then joined to the U.S. Census defined Incorporated Places and Census Designated Places (CDPs) in Texas and Louisiana. Due to a few duplicate place names in Texas (e.g., two municipalities named "Mesquite, Texas"), the Twitter coordinates of tweets with these place names were checked against the actual locations of both places. Subsequently, the dependent variables were manually updated to correct any errors.

After preparing the data, it was apparent that #AllLivesMatter and #TCOT both received considerable attention during the study period. However, a rough analysis of the text of both counter-narratives revealed that #AllLivesMatter tweets far more often directly engage with race-related topics. Therefore, #AllLivesMatter is used as the predominant counter-narrative to #BlackLivesMatter in this study. The two text and profile variables are normalized by population and multiplied by 1000 for readability. The following four Twitter variables serve

as dependent variables (DVs):

- Twitter users per 1000 people with ‘BlackLivesMatter’ in the text of a tweet (BLMTWEPERCAP),
- Twitter users per 1000 people with ‘AllLivesMatter’ in the text of a tweet (ALMTWEPERCAP),
- Twitter users per 1000 people with ‘BlackLivesMatter’ in their profile description (BLMPROFPERCAP), and
- Twitter users per 1000 people with ‘AllLivesMatter in their profile description (ALMPROFPERCAP).

Two datasets from the U.S. Census are used in tandem with Twitter data: the 2014 American Community Survey (ACS) 5-year estimates of Selected Characteristics of the Native and Foreign Born Populations (SO501) as well as Selected Economic Characteristics (DP03) for U.S. Census-defined Incorporated Places and CDPs. These variables include:

- Population (POP)
- Median age (MEDAGE)
- Percent white (PERWHITE)
- Median family income (MEDFAMINC)
- Percent unemployed (PERUNEMP)

A dummy variable representing the state (ST; with Louisiana = 1), is used to test for rough regional effects. Percent black, while a seemingly more logical variable to use in studying #BlackLivesMatter, was neglected in favor of PERWHITE, which captures the presence (or absence) of other racial groups. Multiple racial variables were not used due to concerns over multicollinearity.

Most of the raw variables are highly skewed, significantly deviating from the normal

distribution. Due to this, transformations to several variables were implemented prior to the regression analysis. Numerous transformations, and combinations of transformations, were tested for use in OLS. Additionally, other non-linear regression models were explored as alternatives. However, due to a desire for parsimony, ease of interpretation, and the desire to carry out comparisons between models, OLS with a systematic set of transformations was used in every model. While some procedures, such as Box-Cox, iteratively determine the best transformation for each DV, the resulting disparate transformations would make comparison between models impossible.

A cube root transformation is applied to the dependent variable in each model. This transformation has several advantages over others: it allows for zero values and simultaneously poses no problems for normalized variables, unlike a log transformation. POP, MEDFAMINC, and PERUNEMP, on the other hand, are log transformed. It was expected that MEDAGE would be positively associated with the four dependent variables since younger populations have the greatest rates of LBSM usage (Greenwood et al. 2016). It was expected that PERWHITE would be negatively associated with #BlackLivesMatter DVs and positively associated with the #AllLivesMatter DVs. MEDFAMINC was expected to be negatively associated with the #BlackLivesMatter DVs and positively associated with the #AllLivesMatter dependent variables, since blacks have historically had lower incomes (Massey and Denton 1998). PERUNEMP is included in addition to MEDFAMINC because median income is not a flawless measure of affluence and could potentially mask economic problems leading to racial unrest. A place can have a large median income relative to its unemployment rate as in the case of South Africa, one of the most disparate countries in the world (Tregenna 2011), and under-employment has been a common protest topic dating back to protests inspired by Dr. Martin Luther King, Jr. (Jones 2013).

It was expected that POP would be positively associated with each DV, since other studies have revealed a disproportionate Twitter presence in highly populated places, even after normalizing by population (see Haffner 2018, Hecht and Stephens 2014). However, the DVs

with the highest values have some of the lowest POP values in the dataset, since these municipalities need only one or two users to return proportionally large values. Therefore, municipalities with populations less than 3000 were excluded from the analysis. Though this breakpoint is admittedly arbitrary, it produces a healthy mix of high and low population municipalities with large dependent variable values. After this, observations with missing data and zero values for PERUNEMP were eliminated, resulting in a final sample size of 704 places.

Visualization is carried out in a manner similar to Graham and Zook (2011) and Fekete (2015), noting the most interesting cases, illuminating patterns of spatial similarity and difference, discussing deviations from expected patterns, and individually examining outliers. As previously stated, regression techniques are used in an exploratory fashion, not inferentially. For this reason, regression results must be interpreted with caution. First, the census variables used are residential, and there is a disconnect between these variables and Twitter activity (Haffner 2018; Longley et al. 2015) because users can tweet from locations where they do not live. The fallacy here is potentially ascribing the wrong characteristics to users. Further, since users can select a place from a list of locations when using general location on Twitter, the census variables may not be attributable to the user's actual location, or the closest location, at the time of a post. Users are also mobile, able to travel and tweet from multiple locations.

The approach in this study is novel in that it utilizes information in users' profile descriptions in addition to tweet text, and it makes use of general location (i.e. city) in addition to precise location. Though precise location is usually more desirable because of its specificity, a far greater number of Twitter users opt to use general location, resulting in a larger dataset. The number of geolocated Twitter objects containing the desired texts with precise location was indeed very small compared to those utilizing general location (roughly 6%). In the next section, I begin by discussing broad trends in the data, and then I review regression assumptions and results. As a follow up to regression analysis, I investigate temporal differences between users in Texas and Louisiana during the month of Alton Sterling's death, spatial patterns in the raw variables and residuals, and characteristics of several outliers. I conclude the article with a

review of the major findings and broad implications for future research with such data.

Results and Discussion

Overall, #BlackLivesMatter (6695 users) received much greater attention than #AllLivesMatter (2382 users) in the text of tweets during the study period (Table 1). Similarly, a far greater number of users mentioned #BlackLivesMatter (2702 users) than #AllLivesMatter (127 users) in their Twitter profiles. It is significant to note that the proportion of the number of users mentioning each phrase in a profile description to the number of users mentioning each phrase in the text of a tweet is much larger for #BlackLivesMatter (40.0%, versus 5.3% for #AllLivesMatter). In the context of color-blind racism, this is consistent: few users are willing to bear the banner of the oppositional #AllLivesMatter phrase in their Twitter profile, a reasonable number are willing to use the phrase in the text of tweets, and many more people likely harbor prejudice views offline without discussing the phrases on Twitter at all.

Regression Analysis

In general, the models hold up to regression assumptions fairly well, with the #BlackLivesMatter models performing better than the #AllLivesMatter models. Variance inflation factors are low, with values of 1.186, 1.459, 1.405, 1.535, 1.266, and 1.336 for log(POP), MEDAGE, PERWHITE, log(MEDFAMINC), log(PERUNEMP), and ST, respectively. The models using BLMTWEPERCAP and BLMPROFPERCAP as DVs demonstrate homoscedasticity in their residuals as found through the Breusch-Pagan test. The models using ALMTWEPERCAP and ALMPROFPERCAP as DVs fail to meet this assumption. Similarly, the histograms of residuals in the BLMTWEPERCAP and BLMPROFPERCAP (Fig. 2) regression models appear relatively normally distributed while ALMTWEPERCAP and ALMPROFPERCAP tend towards negatively skewed. The plots of residual versus fitted values each exhibit a straight line trend, which upon investigation is produced by the many zero values in the dataset. These are, however, relevant data points that

should not be excluded, since a lack of #BlackLivesMatter and #AllLivesMatter content is meaningful information.

While these diagnostics should invoke caution against using the models predicatively, the comparisons between models are quite informative and reveal many intriguing results (Tables 2 - 5). Each model has at least two significant variables, namely log(POP) and log(MEDFAMINC), and both have a positive effect on the four DVs. The model using BLMTWEPERCAP as a DV returns the greatest R-squared value at 0.267. This is followed by ALMTWEPERCAP, BLMPROFPERCAP, and ALMPROFPERCAP with R-squared values of 0.237, 0.226, and 0.119, respectively.

In the BLMTWEPERCAP model, every variable is statistically significant except for log(PERUNEMP). The strongest variables are log(POP) and log(MEDFAMINC), with *t*-values of 8.598 and 7.118, respectively. These are followed by PERWHITE which has a *t*-value of -5.198. As expected, the effect of PERWHITE is negative, yet its effect is not the strongest; the effects of log(POP) and log(MEDFAMINC) are stronger. On the other hand, in the BLMPROFPERCAP model, the most significant variables are log(MEDFAMINC) and PERWHITE, with *t*-values of 9.174 and -6.610, respectively.

The variable log(POP) has a stronger effect in the #AllLivesMatter models than in the #BlackLivesMatter models. It has a *t*-value of 8.598 in the BLMTWEPERCAP model but a *t*-value of 10.827 in the ALMTWEPERCAP model. Similarly, it has a *t*-value of 3.493 for BLMPROFPERCAP but a *t*-value of 7.357 for ALMPROFPERCAP. This indicates that the placement of #AllLivesMatter in a profile is more driven by population, noting #BlackLivesMatter's greater reach to places with smaller populations.

The effect of PERWHITE is negatively significant in every model except for ALMPROFPERCAP (t = -1.669). While it was expected that PERWHITE would have a negative effect on the #BlackLivesMatter DVs, its negative effect on ALMTWEPERCAP is counterintuitive. Yet, the effect of PERWHITE is much more significant for both

#BlackLivesMatter variables, indicating that race does play a notable role in the production of such content. The effect of log(MEDFAMINC) is positively significant in every model. In the context of the coefficients of PERWHITE, this is counterintuitive. In each model, the effect of PERWHITE is negative while the effect of log(MEDFAMINC) is positive. Given the historic positive relationship between income and white populations (Massey and Denton 1998), it may be expected that the coefficients of these variables would have the same sign. While log(MEDFAMINC) and PERWHITE have a positive Pearson's *r* (0.21), the VIFs of both variables are low, implying a lack of an interaction between the two. That is, it cannot be assumed that wealthy non-white populations are the primary producers of #BlackLivesMatter and #AllLivesMatter content. Using aggregated spatial data, especially at this coarse spatial resolution, makes such an issue difficult to untangle. That said, income affects rates of contribution to Twitter more generally (Greenwood et al. 2016), so the effects of race and income in this study may be independent.

Differences between Louisiana and Texas users in July 2016

The dummy variable representing the state has a significant, positive effect on both #BlackLivesMatter DVs. While the average BLMTWEPERCAP value in Texas is 0.183, it is nearly twice as large in Louisiana at 0.309. Similarly, the average BLMPROFPERCAP value is 0.421 in Texas and 0.629 in Louisiana. Given the magnitude and location of Alton Sterling's death in July of 2016 (Baton Rouge), this is unsurprising. A closer look at the Twitter activity during this particular time, however, reveals a more detailed picture. The number of users referencing #BlackLivesMatter and #AllLivesMatter peaks in both Texas and Louisiana around July 7 - 8, several days after Alton Sterling's death (Fig. 3). In both states, the #BlackLivesMatter content increases at a faster rate (i.e., the line has a greater slope). Yet, the #BlackLivesMatter references in Louisiana increase faster following July 5, and tail off slower after its peak. At the same time, the proportion of #AllLivesMatter content is much greater in Texas, coinciding with the deaths of the five Dallas police officers who were killed in the wake of Alton Sterling's death.

In general, trends of the two phrases correspond quite well (in both states), with #BlackLivesMatter references slightly preceding those of #AllLivesMatter. The discourse of the two phrases indeed appears to be a conversation in many tweets. Of the tweets that mention #AllLivesMatter, nearly 20% also reference #BlackLivesMatter in the same tweet. In some cases when both phrases are mentioned, it is quite clear what the user is trying to say as in the following:

- "[sic] #AllLivesMatter , will be true once #BlackLivesMatter is included …"

Other users summed it up this way:

- "All lives matter is only insulting when it's used in response to #blacklivesmatter. But in reality all lives do matter."
- "#BlackLivesMatter does not mean we Only Matter or mean more or less than #AllLivesMatter or #BlueLivesMatter it means we Matter Too."

Of course, the converse exists as well, with a clear lack of support for #BlackLivesMatter:

- "Yes, #blacklivesmatter but in the eyes of the non-arrogant #ALLLIVESMATTER."

A non-trivial number of these tweets simply read "#BlackLivesMatter #AllLivesMatter" (or vice versa), making it difficult to determine the user's stance. While machine learning techniques could contribute to untangling questions of attention versus support, these circumstances combined with the difficulty of detecting sarcasm (Bharti et al. 2016) speak to the necessity of mixed methods when working with big spatial data (Shelton et al. 2014).

A subtle, yet still apparent, second relative maximum is visible in the graph of Louisiana around July 17. An examination of individual tweets around this date reveals that many are related to a lesser known shooting of three police officers in Baton Rouge that occurred on July 17. This peak is non-existent in Texas. While the number of #BlackLivesMatter tweets in Texas mentioning Alton Sterling's death speaks to the pervasiveness of race-related (social) media across space, the relative maximum in Louisiana

speaks to the local nature of other digital content production. Though beyond the scope of this paper, a closer examination of the Twitter content in the cities where these events occurred – Dallas and Baton Rouge – might reveal more detailed patterns.

Spatial Patterns

Mapping reveals several alluring spatial patterns, and a follow-up analysis sheds light on these. Houston, Texas presents some of the more disparate regional patterns, particularly in the raw values of BLMTWEPERCAP (Fig. 4). In general, there are larger values in the southwest and smaller values toward the east. This also holds true for BLMPROFPERCAP (Fig. 5) and is largely a function of residential racial patterns. Two exceptionally large values stand out: Prairie View (northwest of Houston) and Oak Ridge North (north of Houston). Incidentally, these places have vastly different independent variable characteristics. Prairie View ranks first in the entire dataset by BLMTWEPERCAP and Oak Ridge North ranks fourth (Table 6), but Prairie View has a relatively small percent white population (18.5%) while Oak Ridge North's is quite large (88.2%). Similarly, Prairie View has a very low MEDAGE at 19.9, while Oak Ridge North has one of the larger MEDAGE values in the dataset at 47.3.

As revealed in the map of BLMTWEPERCAP residuals (Fig. 6), these two places also possess some of the largest residuals in the BLMTWEPERCAP model. While the reason for Oak Ridge North's prominence is puzzling, Prairie View's makes sense given its previously mentioned residential characteristics and the presence of a historically black academic institution, Prairie View A&M University (PVAMU). At the same time, the city itself has a relatively small total population which works to inflate BLMTWEPERCAP, a normalized variable. An examination of the content of the tweets in Prairie View reveals many references to locations, both near and far. Three users simultaneously referenced #BlackLivesMatter and the university with the hashtag #PVAMU. Three users referenced PVAMU in tandem with a reference to the University of Missouri (MU), where a series of racial protests started in the fall of 2015 (Fortunato et al. 2017), with tweets such as:

- "#PrayersforMizzou #BlackLivesMatter #PVAMU19 [sic]"
- "#ConcernedStudent1950 #BlackLivesMatter #PVStandswithMizzou #PrayForMizzou [sic]"

One user referenced Dallas on July 8 (following the shooting of Dallas police officers) saying:

- "#Dallas #BlackLivesMatter #propaganda #PrayForPeace [sic]"

The #BlackLivesMatter content in Oak Ridge North is somewhat different. No users reference MU, but two users reference #PhilandoCastille (who was killed in the Saint Paul, Minnesota area) and #AltonSterling (who was killed in Baton Rouge). Further, one user references events that occurred in New York.

The connection between PVAMU and MU through geotagged social media content speaks to the networked nature of places (Wilken 2008). Locations are defined not only by their local characteristics but by their relationships with other places. The apparent absence of a relationship in this dataset between Oak Ridge North, Texas and Columbia, Missouri is likely a function of Oak Ridge's small population, small black population, and its lack of a university. These are excellent examples of offline processes being reflected through the GeoWeb. At the same time, social media content enables a new type of relationship between Prairie View, Texas and Columbia, Missouri. It allows people to identify with other users and places in real time and broadcast this relationship to anyone with an internet connection. In this way, #BlackLivesMatter on Twitter operates as type of "networked public" (body 2010).

In this light, what drives an individual to reference their location (or university) when discussing political issues on social media? In the case of users in Prairie View referencing support for MU, it appears to be a way of showing solidarity with victims of injustice. But why do users choose to explicitly tag their location in such situations? Wilson (2012) suggests that the tagging of location on social media content is a type of "conspicuous mobility," in which users tag their location in places where they want to be seen and at times when they want to be seen. In line with this, a study of college students reveals that most say they prefer to tag their

location when vacationing or visiting a unique location (Haffner et al. 2017). Yet the tagging of a general location (i.e. city) in the wake of racially charged events appears different. Since this type of social media content coincides with external events, users seem to choose whether or not to tag their location and bring attention to an issue within a relatively short period of time. Thus, such content is event-driven first and location-driven second. While it is possible to select a general location where a user is not physically present, it is not likely that users would visit an exotic location solely for the sake of location tagging in the wake of race-related event. Complicating interpretation, the location type used in this study is not as conspicuous as that to which Wilson (2012) refers (e.g., check-in locations, such as restaurants); it is more general and ambiguous.

How people choose to geotag places is a salient, related issue. As stated previously, users can tag any location they choose. Do users simply select the first location suggested by Twitter based on their current location? Or do they choose to tag the place that they most closely identify with? In isolated places there are fewer general location options, but in large metropolitan areas, this is a greater concern: do residents of suburbs choose to tag their exact municipality of residence or the central city that their metropolitan area is a part of? Or do users choose to tag the locations where events occurred instead of their actual location? Surveys and interviews could shed light on this and thereby reduce the "black box" nature of the geotagging process itself. This predicament, while challenging the assumption that people are present in the location they tag, suggests that LBSM could offer new ways of studying how people identify with places.

The text (BLMTWEPERCAP and ALMTWEPERCAP) and profile (BLMPROFPERCAP AND ALMPROFPERCAP) variables undoubtedly measure different processes. The text variables measure the number of users that referenced #BlackLivesMatter or #AllLivesMatter in the text of a tweet in a given location. Conversely, the profile variables are counts of all the geotagged tweets from users who referenced the phrase in a Twitter profile. Under both metrics, users can be counted in multiple cities. However, the number of cities per

user is relatively low for both of the text variables, with users referencing #BlackLivesMatter averaging 1.08 and users referencing #AllLivesMatter averaging 1.03. In fact, 93.7% of users who referenced #BlackLivesMatter and 96.9% of users who referenced #AllLivesMatter in the text of a tweet did so from only one place. Contrast this with the profile variables, where the number of places per user is 3.52 for #BlackLivesMatter and 2.69 for #AllLivesMatter. Further, 36.5% of #BlacklivesMatter users (and 42.5% of #AllLivesMatter users) geotagged a tweet in only one city. Thus, the profile variables in regression models (BLMPROFPERCAP and ALMPROFPERCAP) reflect users' mobile nature more so than the text variables. Given that most users only tweeted about the text "#BlackLivesMatter" or "#AllLivesMatter" within one city, it is tempting to assume that these are better indicators of users' home locations. This should be avoided since users could be geotagging tweets from another location, such as work location, even if they tweet there often.

In general, the residuals appear relatively uncorrelated spatially (see Fig. 6 and 7). From a modeling perspective, this is a positive result since spatial autocorrelation of residuals violates the assumption of independence. Besides this, explicitly spatial models, such as spatially autoregressive models and geographically weighted regression, were not pursued for multiple reasons. First, the size and distribution of places varies greatly throughout both states. Some places cover tens of square kilometres while others are only a fraction of this size. Both states also possess large swaths of areas with populations below the threshold of 3000, leaving many empty and isolated areas. A more critical issue is how to represent these places spatially, since the calculations for these models require each observation to have a singular coordinate pair. While the centroid of units is commonly used for this, large metropolitan areas contain many interweaving cities, and the varied distribution of users throughout these large areas makes reducing them to a singular point rather defeating.

Challenges and future directions

The spatial approach in this article is limited in that it mostly focuses on broad regional

differences between states. While some intra-urban differences are noted, such as those near and around Houston, the choice to use general location of geotagged tweets restricts the scale of analysis. Additionally, many generally assumed differences between U.S. states are largely superficial, so caution must be taken against ecological fallacy. The political classification of some states as "red" and others as "blue", for instance, erodes when examining voting patterns at the county level (Weichelt 2018). Many functional regions ignore state boundaries; a city in one state can easily be more connected culturally, politically, and economically to another state. That said, the effect of the state is significant in both #BlackLivesMatter models, and the coarse spatial resolution used in this study provides a much larger dataset than what is available through those tweets with precise location.

Stefanidis et al. (2013) suggest methods for more precisely determining users' locations through indirect references to location in social media content, termed ambient geospatial information. While such methods could help in determining users' home locations, and therefore more accurately align with conventional demographic characteristics, such an approach would invariably invoke serious privacy concerns. The #BlackLivesMatter movement and its participants remain relatively anonymous.

Very few users reference #AllLivesMatter in their Twitter profile. Considering that Twitter's U.S. racial makeup closely aligns with the general U.S. population (Pew Research Center 2018), the dearth of #AllLivesMatter appearances in Twitter profiles is not likely the result of simply fewer white users. The prominence of #TCOT as found by Ray et al. (2017) is notable, and counts of users referencing this phrase in their Twitter profile could produce better performing models. #TCOT is even more obscure than #AllLivesMatter, and may reveal different spatial patterns. It seemingly ignores race altogether rather than downplay its effect. In this light, the placement of #AllLivesMatter in a Twitter profile may be considered too forward by those stricken with colorblindness, almost to the point of appearing overtly racist.

Conclusion

The regression analyses in this article reveal many significant relationships. log(POP) and log(MEDFAMINC) are positively significant in every model, meaning that places with large populations and large median family incomes produce more Twitter content on #BlackLivesMatter and #AllLivesMatter. PERWHITE has a strong negative relationship with both #BlackLivesMatter DVs. While PERWHITE's relationship with the #AllLivesMatter DVs is also negative, the effect is much stronger in the #BlackLivesMatter models. Race drives #BlackLivesMatter content more, yet income (and thus, potential access to electronic devices) and population are important drivers as well. #BlackLivesMatter content, especially in Houston, is consistent with residential racial patterns.

The post hoc analyses shed light on several important facets of the data. An investigation of the #BlackLivesMatter and #AllLivesMatter content by day in July 2016 reveals subtle differences between Louisiana and Texas, but also brings out the occurrence of a lesser known event in Baton Rouge that occurred in mid-July. An individual examination of several outliers elicits several notable relationships in the data. One outlier, Prairie View, Texas, demonstrates a connection between a historically black university, PVAMU, and MU, which is not found in other towns lacking a university. Broadly evaluating individual tweets in which both #BlackLivesMatter and #AllLivesMatter are mentioned highlights the difficulty in disentangling attention, support, sympathy, and sarcasm. These subtleties speak to the importance of mixed methods in this kind of work and should caution against completely automated methods in analyzing geolocated social media data.

Methodologically, this article suggests a novel method of using information in users' profile descriptions rather than focusing solely on the text content in tweets. Extensions on these methods could help in untangling the differences between support and attention in social media content. The #AllLivesMatter counter-protest content on Twitter aligns with color-blind racial theory quite well. It has received considerably less attention than #BlackLivesMatter on Twitter, and few users are willing to place the phrase in their profile. Despite this, many who neglect to post about #AllLivesMatter likely still ascribe to race-minimizing political policies

and voting patterns. Clearly, attention on social media is not reflective of general public opinion nor does it correlate with a reduction in inequality. With the broad reach of social media and mobile electronics in today's world, research is now necessary on how attention on social media can be used to effectively bring about social change.

```
"place": {
  "id": "18810aa5b43e76c7",
  "url": "https://api.twitter.com/1.1/geo/id/18810aa5b43e76c7.json",
  "place_type": "city",
  "name": "Dallas",
  "full_name": "Dallas, TX",
  "country_code": "US",
  "country": "United States",
  "bounding_box": {
    "type": "Polygon",
    "coordinates": [
      [
        [
          -96.977527,
          32.620678
        ],
        [
          -96.977527,
          33.019039
        ],
        [
          -96.54598,
          33.019039
        ],
        [
          -96.54598,
          32.620678
        ]
      ]
    ]
  },
  "attributes": {}
},
```


Figure 4.1 Example of the place object in a geotagged tweet

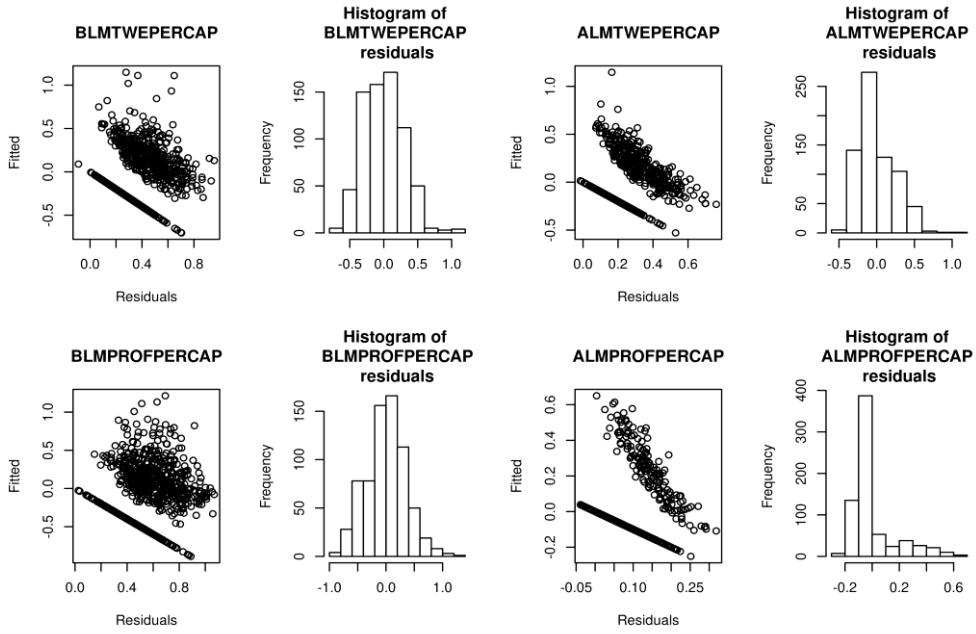

Figure 4.2 Model diagnostics

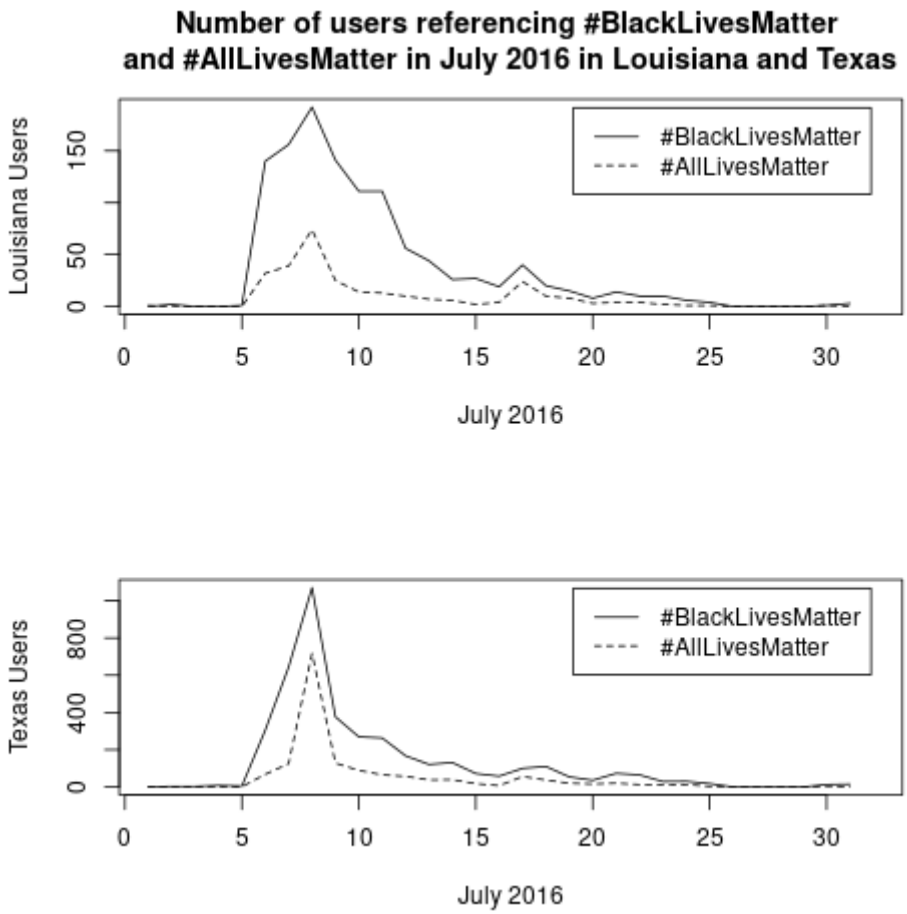

Figure 4.3 Number of users referencing #BlackLivesMatter and #AllLivesMatter in July 2016 in Louisiana and Texas

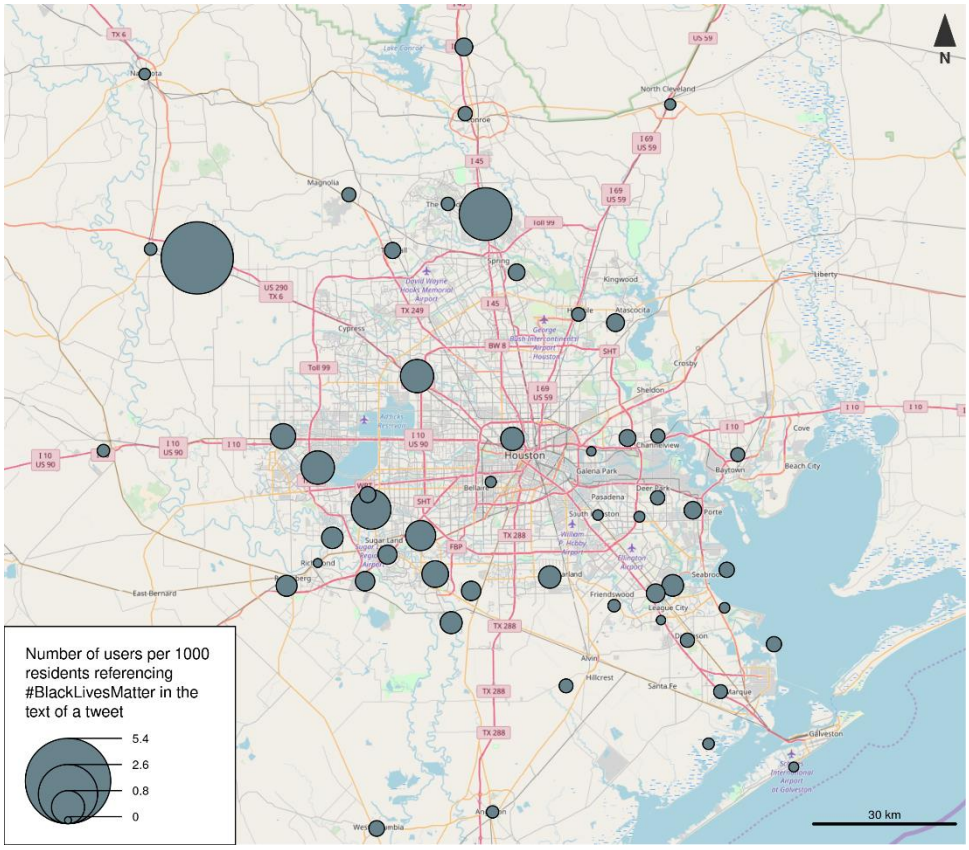

Figure 4.4 Number of users per 1000 residents referencing #BlackLivesMatter in the text of a tweet (BLMTWEPERCAP) in the Houston area

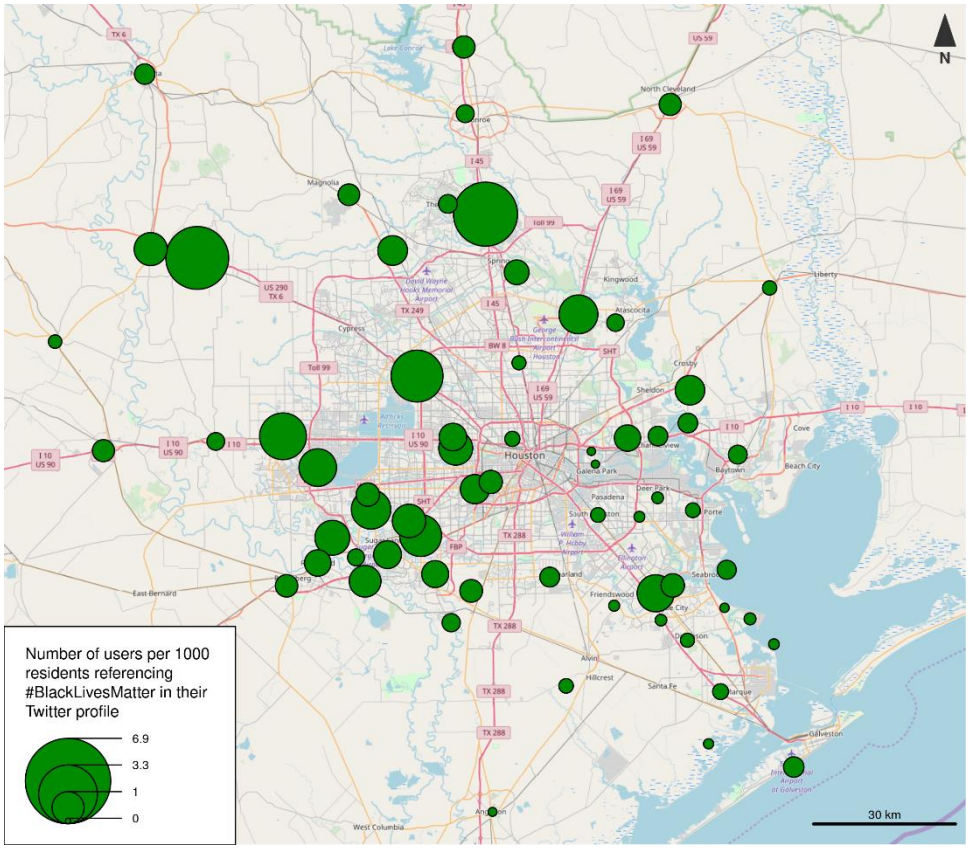

Figure 4.5 Number of users per 1000 residents referencing #BlackLivesMatter in their Twitter profile (BLMPROFPERCAP) in the Houston area

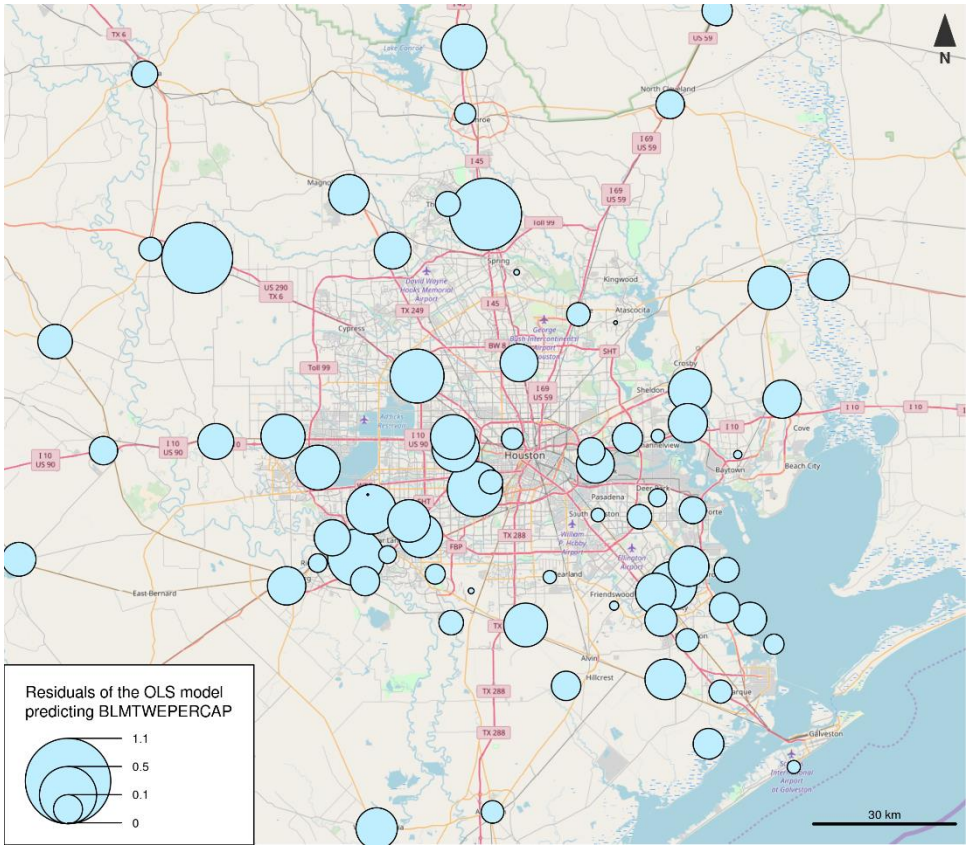

Figure 4.6 Residuals of the OLS model using BLMTWEPERCAP as a DV in the Houston area

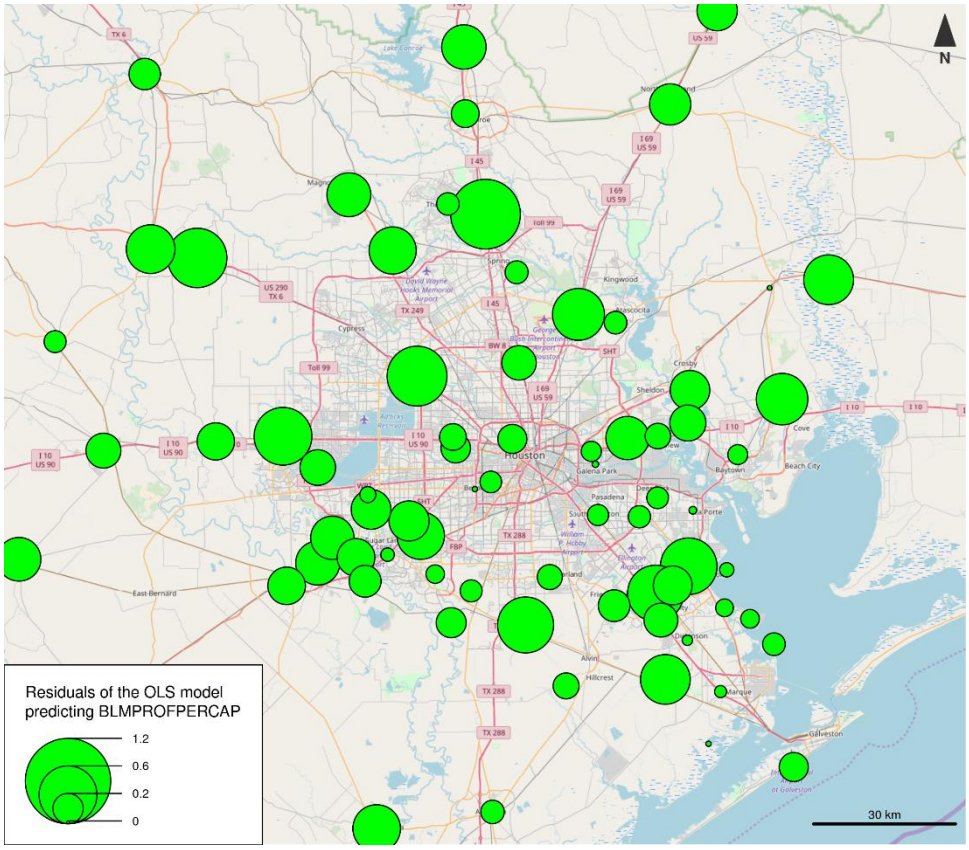

Figure 4.7 Residuals of the OLS model using BLMPROFPERCAP as a DV in the Houston area

Table 4.1 Number of users by phrase and type

	Text of a tweet	Profile
#BlackLivesMatter	6695	2702
#AllLivesMatter	2382	127

Table 4.2 BLMTWEPERCAP Regression Results

	Estimate	Std. Error	T-value	P-value	Sig.
(Intercept)	-2.619	0.328	-7.977	0.000	***
log(POP)	0.094	0.011	8.598	0.000	***
MEDAGE	-0.005	0.002	-2.119	0.034	*
PERWHITE	-0.004	0.001	-5.198	0.000	***
log(MEDFAMINC)	0.230	0.032	7.118	0.000	***
log(PERUNEMP)	0.032	0.024	1.360	0.174	
ST	0.117	0.031	3.756	0.000	***

R-squared: 0.266	Adjusted R-squared: 0.260				
Significance codes:	P<0.001***	P<0.01**	P<0.05*		

Table 4.3 ALMTWEPERCAP Regression Results

	Estimate	Std. Error	T-value	P-value	Sig.
(Intercept)	-1.946	0.264	-7.376	0.000	***
log(POP)	0.095	0.009	10.827	0.000	***
MEDAGE	-0.002	0.002	-0.956	0.340	
PERWHITE	-0.001	0.001	-2.396	0.017	*
log(MEDFAMINC)	0.134	0.026	5.169	0.000	***
log(PERUNEMP)	0.002	0.019	0.103	0.918	
ST	0.034	0.025	1.361	0.174	

R-squared: 0.237	Adjusted R-squared: 0.231				
Significance codes:	P<0.001***	P<0.01**	P<0.05*		

Table 4.4 BLMPROFPERCAP Regression Results

	Estimate	Std. Error	T-value	P-value	Sig.
(Intercept)	-3.355	0.394	-8.521	0.000	***
log(POP)	0.046	0.013	3.493	0.001	***
MEDAGE	0.000	0.003	-0.059	0.953	
PERWHITE	-0.006	0.001	-6.610	0.000	***
log(MEDFAMINC)	0.356	0.039	9.174	0.000	***
log(PERUNEMP)	0.019	0.028	0.670	0.503	
ST	0.082	0.037	2.193	0.029	*

R-squared: 0.226	Adjusted R-squared: 0.220				
Significance codes:	P<0.001***	P<0.01**	P<0.05*		

Table 4.5 ALMPROFPERCAP Regression Results

	Estimate	Std. Error	T-value	P-value	Sig.
(Intercept)	-0.756	0.180	-4.201	0.000	***
log(POP)	0.044	0.006	7.357	0.000	***
MEDAGE	0.000	0.001	-0.141	0.888	
PERWHITE	-0.001	0.000	-1.669	0.096	
log(MEDFAMINC)	0.048	0.018	2.681	0.008	***
log(PERUNEMP)	-0.021	0.013	-1.644	0.101	
ST	0.021	0.017	1.223	0.222	

R-squared: 0.119	Adjusted R-squared: 0.111				
Significance codes:	P<0.001***	P<0.01**	P<0.05*		

Table 4.6 Places with the top 16 BLMTWEPERCAP values and select independent variables

Rank	Place	Users referencing #BLM in the text of a tweet	BLMTWEPERCAP	PERWHITE	PERBLACK	POP	MEDAGE	MEDFAMINC	PERUNEMP
1	Prairie View, TX	32	5.433	18.5	72.8	5890	19.9	64830	7.1
2	Grambling, LA	19	3.789	7.5	90.9	5015	24.6	32241	12.8
3	Commerce, TX	27	3.234	63.3	22.5	8348	23.6	35900	8.8
4	Oak Ridge North, TX	9	2.897	88.2	5.3	3107	47.3	99750	3.6
5	San Marcos, TX	128	2.496	81.1	4.1	51289	23.4	44995	5.9
6	Richland Hills, TX	18	2.269	90	2.3	7933	42.6	63489	6.2
7	Four Corners, TX	23	1.646	22.7	22.5	13973	29.6	62890	3.5
8	Denham Springs, LA	13	1.281	80.4	17.2	10148	36.7	59788	6.8
9	New Orleans, LA	472	1.281	34	59.6	368471	35.1	48381	7.2
10	Baton Rouge, LA	284	1.238	38.7	55	229353	30.8	50119	6.4
11	Cinco Ranch, TX	21	1.165	69.3	6	18028	41.2	144136	2.2
12	Jersey Village, TX	9	1.155	78.1	7.5	7795	40.4	91435	3
13	Ruston, LA	25	1.129	49.3	44.8	22149	24	44880	7.9
14	Old Jefferson, LA	8	1.114	72.2	19.2	7182	34.4	74353	2.6
15	Roanoke, TX	7	1.079	83	2.8	6488	35.1	79375	1.5

Table 4.7 Average number of places per user

	Avg. number of places	St. dev.
BlackLivesMatter text users	1.08	0.32
AllLivesMatter text users	1.03	0.21
BlackLivesMatter profile users	3.52	4.09
AllLivesMatter profile users	2.69	2.57

CHAPTER V

CONCLUSION

Key Findings

This dissertation assessed the validity of LBSM in the study of spatial processes in several different ways. As pointed out by the American Educational Research Foundation (2014) – where validity is a critical concept – validity is not a binary, inherent property of data on its own; validity requires context. The context evaluated in this dissertation, spatial processes, is too broad to provide a definitive answer on whether or not LBSM is valid within it. Rather, subsets of LBSM data are valid in particular contexts that were fleshed out in Chapters II - IV. Chapter II confirms that university students are heavy users of LBSM, and that users often tag content at places they feel are new and exotic, such as vacation spots. In line with this, Chapter III details how the census tracts with the greatest number of users (outliers) possess a significant amenity such as George Bush International Airport, Eleanor Tinsley Park, and Space Center Houston. Using LBSM data to study popular locations within a city, with particular emphasis on younger or university students, is an appropriate context for LBSM data. Similarly, Chapter IV shows that using LBSM to study racial attitudes over broad regions (e.g., states) and across metropolitan regions is also effective.

Chapter II further demonstrates that LBSM, particularly Twitter, is fairly representative of the university demographically with some exceptions. Females and underclassmen are the most likely contributors. There are few racial differences in the perceptions and behaviors of LBSM use – especially with Twitter – suggesting that using LBSM to study race and ethnicity is valid, and

may be able to provide insights on vulnerable populations. Beyond this, students seem to actively engage with place as they post location-enabled content. Place is an important concept, used to further emphasize social media posts. Students' interest in geotagging combined with the wealth of social media platforms that now support location tagging make studying LBSM ever important.

In general, Chapter III provides a more cautionary outlook. The average number of tracts per non-English Twitter user is 2.1, and 68.7% of users only geotagged from one tract during the study period. This demonstrates that most people do not use precise location from very many places, and the survey in Chapter II confirms that few people use precise location in geotagging. Chapter III also uncovers a significant amount of non-stationarity among several independent variables in regression models, especially in PERWHITE. This suggests that individual locations with significant amenities drives geotagging more than the residential characteristics of a place, confirmed through analyzing outlier tracts individually. Precisely geotagged tweets say more about the digital status of locations than the urban dynamics of users. GWR results in other cities would likely yield a different set of significant, and non-stationary, variables. This should invoke caution against using Twitter data in studying day-to-day travel patterns and, more generally, individuals' daily lives. Despite this, several compelling trends are uncovered, such as the anomalous abundance of Turkish users. The number of insignificant statistical relationships is similarly informative, as it suggests that Twitter data captures something that census data does not.

The results of Chapter IV are perhaps the most intriguing in this dissertation. Here, I find a digital manifestation of color-blind racism but not in the manner expected. While PERWHITE has a negative effect on the #BlackLivesMatter regression models, it also has a negative effect on the #AllLivesMatter models, although not as strong. The study in Chapter IV also uncovers significant differences between users in Louisiana and Texas. Examining trends by state over time reveals the occurrence of a lesser known race-related event in which three police officers were killed in Baton Rouge, exhibiting the utility of spatial and temporal stratification in LBSM data

analysis.

Unexpectedly, the comparative results of Chapters III and IV suggest that from a research perspective, the utilization of general location may be more useful than precise location. In Chapter IV, the spatial patterns appear more closely aligned with expected offline processes, the regression models perform better, and there are fewer inexplicable artifacts in the data. This has several possible causes. First, it may simply be due to the relative abundance of tweets geotagged with general location compared to precise location. Alternatively, it may be due to the specific topics pursued. Studying topics other than racial protest and counter-protest with general location may yield poorer results, or other applications, such as studying relationships between users and places, may yet be effective with precise location. Finally, the inexplicable patterns found in Chapter IV, like the large number of Turkish users, may simply be providing new geographic knowledge that is yet to be ground-truthed.

Despite the notable empirical findings in Chapters III and IV, the greater contributions of these chapters are the novel approaches put forth for analyzing LBSM data and the more general insights on the nature of these datasets. The use of profile information, beyond simply tweet text, shows significant promise for LBSM research. The choice to count users rather than tweets appropriately, albeit imperfectly, addresses the long-tail effect, and would be useful in analyses with any social media platform. The subdividing of data by both place and time, as demonstrated in Chapter IV, has the potential to expose new, previously unknown relationships within the data.

Limitations of LBSM Data

This dissertation has uncovered a number of drawbacks to using LBSM as a data source along with hurdles to verifying LBSM's validity in spatial contexts. A significant challenge still remains in eliminating the long-tail effect. Counting the number of users within spatial units reduces, but does not eliminate, the effect of "power users" (Shelton et al. 2015). As stated previously, the average

number of census tracts per user in Chapter III is relatively low, and over half of users only tweeted within one tract. Yet, 2.0% percent of users tweeted from within ten or more tracts and 0.6% tweeted from twenty or more. Thus, this method reduces, but does not eliminate the long-tail effect. The reduction of the influence of power users was greater in Chapter IV however, likely due to the larger spatial unit of analysis. While a primary tweet location could be designated based on where a user tweets most often, assigning individuals to a single location – particular a home location – at a level as fine as census tracts is not feasible. Few users are geotagging with precise location, and the survey results in Chapter II suggest that most users do not geotag content from a singular location, such as their home.

The inability to determine users' home locations also poses challenges for the verification of LBSM data and simultaneously raises privacy concerns if it was effectively addressed. If users' home locations could accurately be determined, their demographic characteristics could be estimated more effectively. Stefanidis et al. (2013) have suggested methods for determining users' locations more accurately, but as briefly discussed in Chapters III and IV, such a process could invoke grave privacy concerns. The potentially vulnerable populations studied in these chapters heightens these concerns. As detailed in a recent first-hand account of cyberharrassment by Cuevas (2018), the internet enables malicious groups to coordinate like never before. With the relative ease of access to data via the Twitter API, a significant challenge exists in developing methods to more effectively use LBSM data while preventing the exploitation of these methods by malevolent radicals. Of course, exploitative practices may be developed independent of academic research, so a more appropriate goal may be developing data accessibility standards that protect users and can be adopted by LBSM platforms, corporations that collect digital data, and greater society. Given that females were discovered to be less concerned about privacy in Chapter II, such issues are critical.

A related but more subtle issue is the state of users' privacy within agencies that collect

digital location data but do not share it. All of the LBSM data in this dissertation were collected from publicly available sources, and thus, the analyses therein mostly pertain to freely available data. However, corporations such as Google, Apple, Facebook and a growing number of mobile application companies have far more detail about users than anything publicly available. Issues such as infrequent references to #BlackLivesMatter in small towns and the lack of non-English Twitter content in some tracts are non-issues for a corporation like Google who collects data on billions of searches everyday across the globe. Conversations about how this data should be used are taking place happening behind closed corporate doors. With relatively lax and unstandardized digital privacy laws, the necessity for a reconceptualization of privacy is paramount (Elwood and Leszcynski 2011).

Despite the encouraging results – and the more ideal approach in terms of users' privacy – of using general location as shown in Chapter IV, this type of location is not specific enough for many applications. Day-to-day urban mobility occurs at a much finer scale than an Incorporated or Census Designated Place, so it is not useful here. Another drawback to LBSM data more generally is the rapidly changing nature of platforms and their users. If Twitter exists in ten years, the facets of how people geotag, users' demographics, and their content production patterns would certainly need revisiting. Further, as other platforms have become popular (e.g., the rise of Instagram in recent years), attention to these is necessary as geographers and data scientists seek to utilize them.

Currently, there is a push within the geographic community to use user-generated content to glean answers in real-time, especially in the context of hazards and disasters (Liu and Palen 2010). Open source projects such as Ushahidi are designed for this purpose and work to combine multiple data sources like social media, blogs, and text messages (Heinzelman and Waters 2010). Beyond hazards and disasters, these projects claim to provide insights for election monitoring, overcrowding, advocacy, and human rights (Ushahidi 2018). Yet, much of this data needs post hoc analysis for verification (Heinzelman and Waters 2010), and data cleaning is a vitally important,

often neglected part of big data analyses (Osborne 2012). The study undertaken in Chapter III required the removal of Twitter "bot" accounts that produce automated content. While the differences in regression results with and without these accounts were negligible, this may not be case in other circumstances. Even after effectively cleaning data, the challenge of identifying patterns remains (Miller and Goodchild 2015). These caveats pose challenges for automated, real-time data analysis.

Many of these drawbacks point to the insufficiencies in using LBSM data on their own. This simultaneously speaks to the importance integrating multiple data sources, the necessity of mixed and qualitative methods, and the value of local knowledge in interpreting big data. A holistic approach to spatial problems with the view of LBSM as a useful tool will be far more beneficial than viewing these new data sources as a cure-all.

Future directions

Despite a number of significant findings, this dissertation has perhaps raised more questions than provided answers. While Chapter II discovered why students geotag content in general, a salient issue pertaining to Chapter IV is why people (not just students) choose to tag their location in the wake of a race-related event, or a social/political event in general. In these cases, why do people choose to geotag content, and how do they perceive their relationship with the place where an event occurred? Pertaining to Chapters III and IV, why do people choose to use precise location versus general location? A more thorough understanding of the psychological processes behind users' geotagging choices would greatly help in making sense of big social media data.

An incredible amount of information exists inside users' Twitter profiles and largely remains untapped. The number of followers and followees, for instance, could function as a useful measure of social connectedness. Combined with users' locations, it could be used to reveal areas that have fewer followers per capita and are thus more socially vulnerable or isolated. Additionally,

the capability of users to reference other users opens a wealth of ways study how people and places are connected. This, of course, would require some locational information but would not require home locations. The varied locations of users combined with the locations of other users they reference in tweets could be used to create a social connectivity matrix of places.

Beyond these applications, there is ripe opportunity for the expansion of LBSM studies in terms of regional focus. The vast majority of such research, especially with Twitter, has been conducted in the United States. This is explicable in part due to publicly available data sources, such as the U.S. Census, which provide important outlets for data quality verification but are not ubiquitous. Yet, the lack of such conventional data sources in other countries may provide opportunities for LBSM to be used as an alternative data source. Though income and internet inequality are serious problems in the United States (Warf 2012), the digital divide may yet be greater in other countries. Thus, digital data sources would be more representative of the elite than other population segments, and extra caution would be required in the interpretation and application of results.

To conclude, LBSM will not be a panacea in solving geographic problems. It will not eliminate the need for conventional data sources, render qualitative and mixed methods unnecessary, or extinguish the vibrancy of theory. It possess demographic biases, measures processes different from conventional data sources, requires special data cleaning procedures, and careful handling of outliers. Nevertheless, LBSM shows promise. Under some contexts it is valid for the analysis of spatial patterns. In this dissertation, it has shown itself useful in identifying a city's popular locations, demonstrated a spatial manifestation of a sociological phenomenon, and elicited new spatial patterns. While a proper understanding of LBSM's drawbacks is crucial, it can be a powerful tool in geographic research.

REFERENCES

Agresti, A. (1990). *Categorical Data Analysis*. New York: John Wiley & Sons.

All My Changes [AMC]. (2015). 6.26: Released April 16, 2015. https://allmychanges.com/p/ios/twitter/ (Accessed 17 November 2016).

Alexander, M. (2010). *The New Jim Crow: Mass incarceration in the age of colorblindness*. New York: New Press.

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired*. 23 June 2008. https://www.wired.com/2008/06/pb-theory/ (Accessed 29 January 2018).

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Dordrecht, Netherlands: Kluwer Academic Publishers.

APKMirror. (2015). Twitter 5.55.0. http://www.apkmirror.com/apk/twitter-inc/twitter/twitter-5-55-0-release/twitter-5-55-0-android-apk-download/ (Accessed 17 November 2016).

Application Delivery Strategies [ADS]. (2001). 3D Data Management: Controlling Data Volume, Velocity, and Variety. https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf (Accessed 12 February 2018).

Baginski, J., Sui, S., & Malecki. E. J. (2014). Exploring the Intraurban Divide Using Online Restaurant Reviews: A Case Study in Franklin County, Ohio. *The Professional Geographer*, *66*, 3, 443–455.

Bartoschek, T., & Kebler. C. (2013). VGI in Education: From K-12 to Graduate Studies. In D. Sui, M. Goodchild, and S. Elwood (Eds.), *Crowdsourcing Geographic Knowledge*, (pp. 341–360). Netherlands: Springer.

Berland, L. (2016). #ThisHappened in 2016. https://blog.twitter.com/official/en_us/a/2016/thishappened-in-2016.html. (Accessed 7 February 2018).

Bharti, S. K., Vachha, B., Pradhan, R. K., Babu, K. S., & Jena, S. K. (2016). Sarcastic sentiment detection in tweets streamed in real time: a big data approach. *Digital Communications and Networks*, *2*, 108–121.

Bivand, R. & Piras, G. (2013). Comparing implementations of estimation models for spatial econometrics. *Journal of Statistical Software*, *45*(2), 150–179.

Bivand, R. & Yu, D. (2017). Spgwr: Geographically weighted regression. R package version 0.6-31. https://CRAN.R-project.org/package=spgwr (Accessed 2 March 2018).

BlackLivesMatter (2018). BlackLivesMatter. https://blacklivesmatter.com (Accessed 14 January 2018).

Bonilla-Silva, E. (2006). *Racism without racists: Color-blind racism and the persistence of racial inequality in America*. 2nd edition. Plymouth, UK: Rowman & Littlefield Publishing Group, Inc.

Box, G. E. P. & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society*, *26*(2), 211–252.

boyd, d. (2010). Social network sites as networked publics: Affordances, dynamics and implications. In Z. Papacharissi (Ed.) *Networked Self: Identity, Community, and Culture on Social Network Sites*, (pp. 39–58). New York: Routledge.

Brandeis, M. W., Zamanillo, C., & Isabel, M. (2017). Finding meaningful participation in volunteer geographic information and citizen science: A case comparison in environmental application. *Cartography and Geographic Information Science, 44*(6), 539–550.

Brandl, S. G., Frank, J., Worden, R. E., & Bynum, T. (1994). Global and specific attitudes toward the police: Disentangling the relationship. *Justice Quarterly, 11*, 119–134.

Brunsdon, C. & Chen, H. (2014). GISTools: Some further GIS capabilities for R. R package version 0.7-4. https://CRAN.r-project.org/package=GISTools (Accessed 2 March 2018).

Brunsdon, C., Fotheringham, A. S., & Charlton, M. E. (1996). Geographically weighted regression: A method for exploring spatial nonstationarity. *Geographical Analysis, 28*(4), 281–298.

Brunson, R. K. (2007). Police don't like black people: African American young men's accumulated police experiences. *Criminology and Public Policy, 11*, 101–132.

Byrd, W. C., Gilbert, K. L., & Richardson Jr., J. B. (2017). The vitality of social media for establishing a research agenda on black lives and the movement. *Ethnic and Racial Studies*, *40*(11), 1872–1881.

Capps, R., Fix, M., & Nwosu, C. (2015). *A profile of immigrants in Houston, the nation's most diverse metropolitan area*. Washington, D.C.: Migration Policy Institute.

Cheshire, J., Barratt, J., Manley, E. & O'Brien, O. (2016). Twitter NYC: A multilingual social city. http://ny.spatial.ly/ (Accessed 1 August 2017).

Cheshire, J., Manley, E. & O'Brien, O. (2016). Twitter tongues: A multilingual social city - the language of tweets in London in summer 2012. http://twitter.mappinglondon.co.uk/ (Accessed 1 August 2017).

Clark, M. (2014). *To tweet our own cause: A mixed-methods study of the online phenomenon "Black Twitter"*. Doctoral dissertation, University of North Carolina at Chapel Hill.

Cramer, H., Rost, M., & Holmquist, L. E. (2011). Performing a Check-in. *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services - MobileHCI '11*. 57–66.

Crampton, J.W., Graham, M., Poorthuis, A., Shelton, T., Stephens, M., Wilson, M. W., & Zook, M. (2013). Beyond the geotag: Situating 'big data' and leveraging the potential of the geoweb. *Cartography and Geographic Information Science, 40*(2), 130–139.

Crutcher, M. & Zook, M. (2009). Placemarks and waterlines: Racialized cyberscapes in post-Katrina Google Earth. *Geoforum, 40*(4), 523–534.

Cuevas, J. (2018). A new reality? The far right's use of cyberharassment against academics. American Association of University Professors. https://www.aaup.org/article/new-reality-far-rights-use-cyberharassment-against-academics#.Wn4ZhEtG3b1 (Accessed 9 February 2018).

de Lange, M. & de Waal, M. (2013). Owning the city: New media and citizen engagement in urban design. *First Monday, 18*(11).

de Souza e Silva, A. (2006). Mobile technologies as interfaces of hybrid spaces. *Space and Culture, 9*(3), 261–273.

DeVan, A. (2016). The 7 V's of big data. Impact Radius Blog. https://www.impactradius.com/blog/7-vs-big-data/ (Accessed 30 January 2018).

Dodge, M. & Kitchin, R. (2005). Code and the Transduction of Space. *Annals of the Association of American Geographers, 95*(1), 162–180.

Donovan, J. (2015). *Technologies of social change: Mapping the infrastructure of the Occupy Movement from #OccupyWallStreet to #OccupySandy*. Doctoral dissertation, University of California, San Diego.

Duggan, M. (2015). Mobile Messaging and Social Media 2015. *Pew Research Center: Internet, Science, & Tech.* Washington, D.C.: Pew Research Center.

Egenhofer, M. J., Clark, K. C., Gao, S., Quesnot, T., Franklin, W. R., Yuan, M., & Coleman, D. (2016). Contributions of GIScience over the past twenty years. In H. Onsrud & W. Kuhn (Eds.) *Advancing geographic information science: The past and next twenty years*, (pp. 9–34). Needham, MA: GSDI Association Press.

Elasticsearch (2017). Elasticsearch BV. https://www.elastic.co/ (Accessed 1 August 2017).

Ellis, M., Wright, R., Holloway, S., & Fiorio, L. (2017). Remaking white residential segregation: metropolitan diversity and residential change in the United States. *Urban Geography*, 1–27.

Elwood, S., Goodchild, M. F., & Sui, D. (2013). Prospects for VGI Research and the emerging fourth paradigm. In D. Sui, S. Elwood, & M. F. Goodchild (Eds.), *Crowdsourcing Geographic Knowledge*, (pp. 361–375). Dordrecht, Netherlands: Springer.

Elwood, S., & Leszcynski, A. (2011). Privacy reconsidered: New representations, data practices, and the geoweb. *Geoforum, 42*, 6–15.

Evans, L. (2015). *Locative Social Media: Place in the Digital Age*. Basingstoke, U.K.: Palgrave-Macmillan.

Farman, J. (2012). *Mobile Interface Theory*. New York: Routledge.

Fekete, E. (2018). Twitter. Forthcoming in B. Warf (Ed.) *Encyclopedia of the Internet*. Sage: Los Angeles.

Fekete, E. (2015). Race and (online) sites of consumption. *Geographical Review*, *105*(4), 472–491.

Fekete, E. & Warf, B. (2013). Information technology and the "Arab Spring". *The Arab World Geographer*, *16*(2), 210–227.

Firican, G. (2017). The 10 Vs of big data. Transforming Data With Intelligence. Upside: Where data means business. https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx (Accessed 30 January 2018).

Flanagin, A. J. & Metzger, M. J. (2008). The credibility of volunteered geographic information. *GeoJournal, 72*(3-4), 137–148.

Fortunato, J. A., Gigliotti, R. A., & Ruben, B. D. (2017). Racial incidents at the University of Missouri: The value of leadership communication and stakeholder relationships. *International Journal of Business Communication*, *54*(2), 199–209.

Frith, J. (2012). Splintered space: Hybrid spaces and differential mobility. *Mobilities*, *7*(1), 131–149.

Goffman, E. (1959). *The Presentation of the Self in Everyday Life*. New York: Doubleday.

Goggin, G. (2012). Encoding place: The Politics of Mobile Location Technologies. In R. Wilken and G. Goggin (Eds.) *Mobile Technology and Place* (pp. 198–212). London: Taylor & Francis.

Goodchild, M. F. (2008). Spatial accuracy 2.0. *Proceedings of the 8th International Symposium on Spatial Accuracy in Natural Resources and Environmental Sciences*, *1*, 1-7.

Goodchild, M. F. (2007a). Citizens as sensors: The world of volunteered geography. *GeoJournal*, *69*(4), 211–221.

Goodchild, M. F. (2007b). Citizens as voluntary sensors: Spatial data infrastructure in the world of Web 2.0. *International Journal of Spatial Data Infrastructures Research 2*, 24–32.

Goodchild, M. F., & Glennon, J. A. (2010). Crowdsourcing geographic information for disaster response: a research frontier. *International Journal of Digital Earth*, *3*(3), 231–241.

Goodchild, M. F., & Li, L. (2012). Assuring the quality of volunteered geographic information. *Spatial Statistics*, *1*, 110–120.

Gordon, E., Baldwin-Philippi, J., & Balestra, M. (2013). *Why We Engage: How Theories of Human Behavior Contribute to Our Understanding of Civic Engagement in a Digital Era*. Cambridge, MA: The Berkman Center for Internet & Society.

Graham, M. & Shelton, T. (2013). Geography and the future of big data, big data and the future of geography. *Dialogues in Human Geography 3*(3), 255–361.

Graham, M., B. Hogan, R. K. Straumann, & Medhat, A. (2014). Uneven geographies of user-generated information: Patterns of increasing information poverty. *Annals of the Association of American Geographers, 104* (4), 746–764.

Graham, M., Hale, S. A., & Gaffney, D. (2014). Where in the world are you? Geolocation and language identification in Twitter. *The Professional Geographer, 66*(4), 568–578.

Graham, R. & Smith, S. (2016). The content of our #Characters: Black Twitter as counterpublic. *Sociology of Race and Ethnicity*, *2*(4), 433–449.

Graham, M. & Zook, M. (2013). Augmented realities and uneven geographies: Exploring the geolinguistic contours of the web. *Environment and Planning A, 45*(1), 77–99.

Graham, M., Zook, M., & Boulton, A. (2012). Augmented Reality in Urban Places: Contested Content and the Duplicity of Code. *Transactions of the Institute of British Geographers, 38*(3), 464–479.

Graham, M. & Zook, M. (2011). Visualizing global cyberscapes: Mapping user-generated placemarks. *Journal of Urban Technology*, *18*(1), 115–132.

Greenwood, S., Perrin, A., & Duggan, M. (2016). *Social media update*. Washington, D.C.: Pew Research Center. http://www.pewinternet.org/2016/11/11/social-media-update-2016/ (Accessed 11 November 2018).

Griffin, G. P. & Jiao, J. (2015). Where does bicycling for health happen? Analysing volunteered geographic information through place and plexus. *Journal of Transport & Health*, *2*(2), 238–247.

Haffner, M. (2018). A spatial analysis of non-English Twitter activity in Houston, Texas (revisions submitted; in review).

Haffner, M., & A. Mathews. (2016). A Multi-Dimensional Topology for Crowdsourced Geographic Information. In *NSF Workshop on Geospatial Data Science in the Era of Big Data and CyberGIS*. Urbana, IL: University of Illinois at Urbana-Champaign. July 25-26.

Haffner, M., Mathews, A. J., Fekete, E., & Finchum, G. A. (2017) Location-based social media behavior and perception: Views of university students. *Geographical Review*. 1–22.

Hafkin, N. J. & Huyer, S. (2006). *Cinderella or Cybrella? Empowering Women in the Knowledge Society.* Boulder, CO: Kumarian Press, Inc.

Haklay, M. & Budhathoki, N. (2010). OpenStreetMap – Overview and Motivational Factors. In *Horizon Infrastructure Challenge Theme Day*. Nottingham: University of Nottingham, UK.

Haklay, M. (2013). Citizen science and volunteered geographic information: Overview and typology of participation. In D. Sui, S. Elwood, & M. F. Goodchild (Eds.), *Crowdsourcing Geographic Knowledge*, (pp. 105–122). Dordrecht, Netherlands: Springer.

Harvey, F. (2013). To volunteer or contribute locational information? Towards truth in labeling for crowdsourced geographic information. In D. Sui, S. Elwood, & M. F. Goodchild (Eds.), *Crowdsourcing Geographic Knowledge*, (pp. 31–42). Dordrecht, Netherlands: Springer.

Havekes, E., Bader, M., & Krysan, M. (2016). Realizing racial and ethnic neighborhood preferences? Exploring the mismatches between what people want, where they search, and where they live. *Population Research and Policy Review*, *35*(1), 101–126.

Hecht, B. & Stephens, M. (2014). A tale of cities: Urban biases in volunteered geographic information. *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014* (pp. 197–205). Ann Arbor, MI: The AAAI Press.

Heinzelman, J. & Waters, C. (2010). Crowdsourcing crisis information in disaster-affected Haiti. United States Institute of Peace. https://www.usip.org/publications/2010/09/crowdsourcing-crisis-information-disaster-affected-haiti (Accessed 11 February 2018).

Horowitz, J. M. & Livingston, G. (2016). How Americans view the Black Lives Matter movement. Pew Research Center. http://www.pewresearch.org/fact-tank/2016/07/08/how-americans-view-the-black-lives-matter-movement/ (Accessed 19 January 2018).

Imi, Y., Hayakawa, T., & Ito, T. (2012). Analyzing the effect of Open Street Map during crises: The Great East Japan Earthquake. *2012 IEEE 14th International Conference on Commerce and Enterprise Computing*. Hangzhou, China.

Ince, J., Rojas, F., & Davis, C. A. (2017). The social media response to Black Lives Matter: How Twitter users interact with Black Lives Matter through hashtag use. *Ethnic and Racial Studies*, *40*(11), 1814–1830.

Institutional Research and Information Management [IRIM]. (2015). *OSU Student Profile: Fall 2015*. Stillwater, OK: Oklahoma State University.

Joiner, R., Stewart, C., & Beaney, C. (2015). *Gender digital divide: Does it exist and what are the explanations?* In L. D. Rosen, N. Cheever, and L. M. Carrier (Eds.), *Wiley Handbook of Psychology, Technology, and Society*. (pp. 74–88). New York: John Wiley & Sons.

Jones, W. P. (2013). *The March on Washington: Jobs, freedom, and the forgotten history of civil rights*. New York: W.W. Norton & Company.

Kay, S., Zhao, B., & Sui. D. (2014). Can Social Media Clear the Air? A Case Study of the Air Pollution Problem in Chinese Cities. *The Professional Geographer, 67*(3), 351–363.

Kennedy, H. (2006). Beyond anonymity, or future directions for internet identity research. *New Media & Society, 8*(6), 859–876.

Kent, J. D. & Capello Jr., H. T. (2013). Spatial patterns and demographic indicators of effective social media content during the Horsethief Canyon fire of 2012. *Cartography and Geographic Information Science, 40*(2), 78–89.

Kitchin, R. (2013). The real-time city? Big data and smart urbanism. *GeoJournal, 79*(1), 1–14.

Lee, J. H., Gao, S., & Goulias, K. G. (2016). *Can Twitter data be used to validate travel demand models?* Washington, D.C.: 95th Annual Transportation Research Board Meeting.

Leetaru, K. H., Wang, S., Cao, G., Padmanabhan, A., & Shook, E. (2013). Mapping the global Twitter heartbeat: The geography of Twitter. *First Monday, 18*(5-6).

Li, L., Goodchild, M. F., & Xu, B. (2013). Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *Cartography and Geographic Information Science, 40*(2), 61–77.

Lindqvist, J., Cranshaw, J., Wiese, J., Hong, J., & Zimmerman, J. (2011). I'm the mayor of my house: Examining why people use Foursquare – a social-driven location sharing application. *Proceedings of the 2011 Annual Conference on Human factors in Computing Systems - CHI '11*, 2409–2418.

Liu, S. B. & Palen, L. (2010). The new cartographers: Crisis map mashups and the emergence of neogeographic practice. *Cartography and Geographic Information Science*, *37*(1), 69–90.

Longley, P. A., Adnan, M., & Lansley, G. (2015). The geotemporal demographics of Twitter usage. *Environment and Planning A: Economy and Space*, *47*(2), 465–484.

Machin, J. (2017). Twitter's most popular tweets and accounts of 2017. https://mashable.com/2017/12/05/twitter-most-popular-2017/#RNrG_p1irqqp (Accessed 7 February 2018).

Malpas, J. (2012). The place of mobility: Technology, connectivity, and individualization. In R. Wilken and G. Goggin (Eds.) *Mobile Technology and Place* (pp. 26–38). London: Taylor & Francis.

Massey, D. & Denton, N. (1998). *American Apartheid: Segregation and the making of the underclass*. Boston, MA: Harvard University Press.

Mathews, A., Lu, Y., Patton, M., Dede-Bamfo, N., & Chen. J. (2013). College students' consumption, contribution, and risk awareness related to online mapping services and social media: Does geography and GIS knowledge matter? *GeoJournal*, *78*(4), 627–639.

McNaught, C. & P. Lam. (2010). Using Wordle as a supplementary research tool. *The Qualitative Report*, *15*(3), 630–643.

Miller, H. J. (2017). Geographic information science II: Mesogeography: Social physics, GIScience and the quest for geographic knowledge. *Progress in Human Geography*, 1–10.

Miller, H. J. (2010). The data avalanche is here. Shouldn't we be digging? *Journal of Regional Science*, *50*(1), 181–201.

Miller, H. J. & Goodchild, M. F. (2015). Data-driven geography. *GeoJournal*, *80*(4), 449–461.

Navratil, G. & Frank, A. U. (2013). VGI for land administration – a quality perspective. *International Archives of the Photogrammetry, Remote Sensing, and Spatial Information Sciences, 42*(1), 1–5.

O'Reilly, T. (2005). What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. http://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html (Accessed 1 August 2017).

Osborne, J. (2012). *Best practices in data cleaning: A complete guide to everything you need to do before and after collecting your data*. Thousand Oaks, CA: Sage Publications, Inc.

Parr, D. (2015). The Production of Volunteered Geographic Information: A Study of OpenStreetMap in the United States. Doctoral dissertation, Texas State University.

Pew Research Center. (2018). Social media fact sheet. http://www.pewinternet.org/fact-sheet/social-media/ (Accessed 7 February 2018).

Rainie, L. & Perrin, A. (2017). 10 facts about smartphones as the iPhone turns 10. Washington, D.C.: Pew Research Center. Retrieved from http://www.pewresearch.org/fact-tank/2017/06/28/10-facts-about-smartphones/ (Accessed 1 August 2017).

Ray, R., Brown, M., Fraistat, N., & Summers, E. (2017). Ferguson and the death of Michael Brown on Twitter: #BlackLivesMatter, #TCOT, and the evolution of collective identities. *Ethnic and Racial Studies, 40*(11), 1797–1813.

Rios, V. M. (2011). *Punished: Policing the Lives of Black and Latino Boys*. New York: New York University Press.

Sax, L., Gilmartin, S., & Bryant, A. (2003). Assessing response rates and nonresponse bias in web and paper surveys. *Research and Higher Education, 44*(4), 409–432.

Schaafsma, S. (2018). Big data: The 6 Vs you need to look at for important insights. Motivation Research and Strategy. https://www.motivaction.nl/en/news/blog/big-data-the-6-vs-you-need-to-look-at-for-important-insights (Accessed 30 January 2018).

Schwartz, R. & Halegoua, G. R. (2015). The Spatial Self: Location-based Identity Performance on Social Media. *New Media & Society*, *17*(10), 1643–1660.

See, L., Mooney, P., Foody, G., Bastin, L., Comber, A., Estima, J., Fritz, S., Kerle, N., Jiang, B., Laakso, M., Liu, H., Milˇcinski, G., Nikšiˇc, M., Painho, M., Pdör, A., Olteanu-Raimond, A., & Rutzinger, M. (2016). Crowdsourcing, citizen science or volunteered geographic information? The current state of crowdsourced geographic information. *ISPRS International Journal of Geo-Information*, *5*(5), 55.

Shafer, T. (2017). The 42 V's of big data and data science. Elder Research: Data Science \& Predictive Analytics. https://www.elderresearch.com/company/blog/42-v-of-big-data (Accessed 30 January 2018).

Shelton, T., Poorthuis, A., & Zook, M. (2015). Social media and the city: Rethinking urban socio-spatial inequality using user-generated geographic information. *Landscape and Urban Planning 142*, 198–211.

Shelton, T., Poorthuis, A., Graham, M., & Zook, M. (2014). Mapping the data shadows of Hurricane Sandy: Uncovering the sociospatial dimensions of big data. *Geoforum*, *52*, 167–179.

Smith, A. (2016). Record shares of Americans now own smartphones, have home broadband. Pew Research Center. http://www.pewresearch.org/fact-tank/2017/01/12/evolution-of-technology/ (Accessed 11 November 2016).

Stefanidis, A., Crooks, A., & Radzikowski, J. (2013). Harvesting ambient geospatial information from social media feeds. *GeoJournal 78*(2), 319–338.

Stephens, M. (2013). Gender and the GeoWeb: Divisions in the production of user-generated cartographic information. *GeoJournal 78*(4), 981–996.

Sui, D. & Goodchild, M. (2011). The convergence of GIS and social media: Challenges for GIScience. *International Journal of Geographic Information Science 25*(11), 1737– 1748.

R Core Team. (2016). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. https://www.R-project.org (Accessed 11 November 2016).

R Core Team. (2017). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. https://www.R-project.org (Accessed 1 August 2017).

Tregenna, F. (2011). Earnings inequality and unemployment in South Africa. *International Review of Applied Economics, 25*(5), 585–598.

Trimble, J. E. & Dickson, R. (2007). Ethnic identity. In C. B. Fisher & R. M. Lerner (Eds.), *Encyclopedia of Applied Developmental Science*. Thousand Oaks, California: Sage Publications. Retrieved from http://sk.sagepub.com/reference/applieddevscience/n160.xml

Twitter. (2017). Twitter developer documentation: Twitter for websites supported languages. https://dev.twitter.com/web/overview/languages (Accessed 1 August 2016).

Twitter. (2016). Adding location to a tweet. https://support.twitter.com/articles/122236 (Accessed 11 November 2017).

Ushahidi. (2018). Ushahidi. https://www.ushahidi.com/ (Accessed 11 February 2018).

Warf, B. (2001). Segueways into Cyberspace: Multiple Geographies of the Digital Divide. *Environment and Planning B: Planning and Design*, *28*(1), 3–19.

Warf, B. (2012). Contemporary digital divides in the United States. *Journal of Economic and Social Geography, 104*(1), 1–17.

Warren, P. Y. (2011). Perceptions of police disrespect during vehicle stops: A race-based analysis. *Crime & Delinquency*, *57*(3), 356–376.

Weichelt, R. (2018). 2016 United States presidential election percent Democrat at the county level. https://people.uwec.edu/weicherd/ (Accessed 16 January 2018).

Weitzer, R. & Tuch, S. A. (2006). *Race and policing in America: Conflict and reform*. Cambridge, UK: Cambridge University Press.

——— (2005). Racially biased policing: Determinants of citizen perceptions. *Social Forces*, *83*(3), 1009–1030.

——— (2004). Race and perceptions of police misconduct. *Social Problems*, *51*(3), 305–325.

Wilken, R. (2008). Mobilizing place: mobile media, peripatetics, and the renegotiation of urban places. *Journal of Urban Technology*, *15*(3), 39–55.

Wilson, M. (2012). Location-based services, conspicuous mobility and the location-aware future. *Geoforum 43*(6), 1266–1275.

Zickuhr, K. (2013). Location-based Services. Pew Research Center. http://www.pewinternet.org/2013/09/12/location-based-services/ (Accessed 11 November 2016).

Zook, M. & M. Graham. (2007). The creative reconstruction of the internet: Google and the privatization of cyberspace and digiplace. *Geoforum, 38*(6), 1322–1343.

Zook, M., Barocas, S., boyd, d., Crawford, K., Keller, E., Gangadharan, S. P., Goodman, A., Hollander, R., Koenig, B. A., Metcalf, J., Narayanan, A., Nelson, A., & Pasquale, F. (2017). Ten simple rules for responsible big data research. *PLOS Computational Biology*, *13*(3), 1–10.

Zook, M., Graham, M., Shelton, T., & Gorman, S. (2010). Volunteered geographic information and crowdsourcing disaster relief: A case study of the Haitian earthquake. *World Medical & Health Policy, 2*(2), 7–33.

VITA

Matthew Haffner

Candidate for the Degree of

Doctor of Philosophy

Thesis: ASSESSING THE VALIDITY OF LOCATION-BASED SOCIAL MEDIA IN THE STUDY OF SPATIAL PROCESSES

Major Field: Geography

Biographical:

Education:

Completed the requirements for the Doctor of Philosophy in Geography at Oklahoma State University, Stillwater, Oklahoma in May, 2018.

Completed the requirements for the Master of Science in Geography at Oklahoma State University, Stillwater, Oklahoma in 2014.

Completed the requirements for the Bachelor of Science in Secondary Education at Pittsburg State University, Pittsburg, Kansas in 2012.

Experience:

Graduate Research Assistant, Department of Geography, Oklahoma State University, Stillwater, Oklahoma (January 2016)
Graduate Teaching Assistant, Department of Geography, Oklahoma State University, Stillwater, Oklahoma (August 2012 – December 2015)
Programmer, Department of Plant and Soil Sciences, Oklahoma State University, Stillwater, Oklahoma (May 2015 – August 2015)

Professional Memberships:

American Geographical Society; American Association of Geographers; Gamma Theta Upsilon; OSU Forum of Geography Graduate Students